Genome sequence of the ornamental plant *Digitalis purpurea* reveals the molecular basis of flower color and morphology variation

Jakob Maximilian Horz[1,2], Katharina Wolff[1], Ronja Friedhoff[2], Boas Pucker[1*]

[1] Plant Biotechnology and Bioinformatics, IZMB, University of Bonn, Kirschallee 1, 53115 Bonn, Germany
[2] Plant Biotechnology and Bioinformatics,TU Braunschweig, Mendelssohnstrasse 4, 38106 Braunschweig, Germany

* corresponding author: pucker@uni-bonn.de

ORCIDs:
JMH: 0009-0007-7367-9439
KW: 0000-0002-3616-0482
RF: 0009-0002-0270-4661
BP: 0000-0002-3321-7471

Abstract

Digitalis purpurea (foxglove) is a widely distributed ornamental plant. Here, we present a long read sequencing-based genome sequence of a magenta flowering *D. purpurea* plant and a corresponding prediction of gene models. The high assembly continuity is indicated by the N50 of 4.3 Mbp and the completeness is supported by discovery of about 96% complete BUSCO genes. This genomic resource paves the way for an in-depth investigation of the flower pigmentation of *D. purpurea*. Structural genes of the anthocyanin biosynthesis and the corresponding transcriptional regulators were identified. The comparison of magenta and white flowering plants revealed a large insertion in the anthocyanidin synthase gene in white flowering plants that most likely renders this gene non-functional and could explain the loss of anthocyanin pigmentation. Furthermore, we found a large insertion in the *DpTFL1/CEN* gene to be likely responsible for the development of large terminal flowers.

Keywords:

Genome sequence, foxglove, flavonoids, anthocyanins, pigmentation, ONT sequencing, MYB

Introduction

Digitalis purpurea, commonly known as foxglove, is a biennial plant in the Plantaginaceae family [1]. The species is native to Europe, but appears in many other parts of the world. *D. purpurea* has been used for centuries for its medicinal properties, particularly for treating heart conditions such as atrial fibrillation and congestive heart failure [2, 3]. The plant contains a variety of cardiac glycosides, with the most important being digoxin [2]. These compounds work by increasing the force and efficiency of heart contractions, which are used to treat heart failure and other cardiac conditions. *D. purpurea* and the close relative *D. lanata* are cultivated for the extraction of cardenolides [1]. Scientists have been studying the plant for many years to better understand the chemistry of the cardiac glycosides it produces and their effects on the human body. Besides its medicinal value, *D. purpurea* is also famous for beautiful flowers (**Fig. 1a**). It is a tall plant that can reach up to six feet in height and produces striking magenta flowers on tall spikes. Various *Digitalis* species display a plethora of diverse flower colors and shapes. While the plant is visually appealing, it is considered toxic due to the formation of specialized metabolites like the cardenolides. Nonetheless, *D. purpurea* is widespread as an ornamental plant. Contributing to the flower colors are anthocyanins, the product of one specific branch of the flavonoid biosynthesis (**Fig. 1b**). The substrate of the flavonoid biosynthesis is the aromatic amino acid phenylalanine, which is channeled through the general phenylpropanoid pathway and the flavonoid biosynthesis towards anthocyanins [4, 5]. Major branches of the flavonoid biosynthesis lead to colorful anthocyanins, colorless flavonols, colorless flavones, and proanthocyanidins which appear dark after oxidation. The anthocyanin biosynthesis shares several enzymatic steps with other pathways like the proanthocyanidin biosynthesis. Dihydroflavonol 4-reductase (DFR), anthocyanidin synthase (ANS), UDP-glucose dependent 3-O-anthocyanidin-glucosyltransferase (UFGT), and other decorating enzymes are often considered as the anthocyanin biosynthesis branch [4, 6] (**Fig. 1c**). A recent study revealed anthocyanin-related glutathione S-transferase (arGST) as an additional enzyme in the anthocyanin biosynthesis [7]. After biosynthesis of anthocyanins at the endoplasmic reticulum, a transport through the cytoplasm and import into the central vacuole is required for long term storage. This process involves several proteins including two tonoplast-localized transporters and a tonoplast-localized ATPase for generation of a proton gradient [5, 8, 9]. A previously postulated ‘ligandin’, that might protect anthocyanins during the transport through the cytoplasm [10], was recently characterized as an enzyme in the cyanidin biosynthesis, which makes an involvement in the anthocyanin transport less likely [7]. Besides direct transport through the cytoplasm, there are several studies that report an involvement of vesicles in the anthocyanin transport in different plant species [5, 11–14].

The pigmentation of a plant structure is not only determined by the activity of the anthocyanin biosynthesis, but also by the activity of pathways that compete for the same substrate. High

activity of the flavonol biosynthesis can deplete the available substrate of the anthocyanin biosynthesis branch [15]. Subtle differences in substrate specificity of the competing enzymes ,e.g. flavonol synthase (FLS) and DFR, can determine the pattern of synthesized flavonoids [16, 17]. Differences in expression of the first committed genes *DFR* and *FLS*, respectively, appear as a decisive factor in controlling the activity of the anthocyanin and flavonol biosynthesis branch [17].

Activity of different branches of the flavonoid biosynthesis is mostly controlled at the transcriptional level through regulation of gene expression by transcription factors [18–21]. Decades of research turned the regulation of the flavonoid biosynthesis into a model system for research on cis- and trans-regulatory elements in eukaryotes. The anthocyanin biosynthesis genes are activated by a complex comprising three different proteins: a R2R3-MYB protein, a bHLH protein, and a WD40 protein [21–23]. Different subgroup 6 MYBs and subgroup IIIf bHLHs can be included in this complex leading to a diverse set of different complexes which differ in their ability to activate certain target genes [24, 25]. The WD40 component is often TTG1, but LWD1 and LWD2 can also participate in complex formation [26]. An additional WRKY component (TTG2) has been suggested to specifically increase the expression of genes associated with the anthocyanin transport into the vacuole [21, 27]. Anthocyanin biosynthesis-regulating MYBs have been reported in different MYB clades. PAP1 (MYB75)/PAP2 (MYB90)/PAP3 (MYB113)/PAP4 (MYB114) are well known anthocyanin regulators in *Arabidopsis thaliana* [23, 28, 29]. While most anthocyanin regulating MYBs are activators, the incorporation of subgroup 4 MYBs into the MBW complex can result in an inhibitory function [30].

Pigmentation differences between plants of the same species are often due to genetic variations affecting the functionality or gene expression of the MYB and bHLH transcription factors [31–34]. This results in extremely low or no transcriptional activation of specific flavonoid biosynthesis genes under certain conditions. However, activation of the structural genes in the context of a different pathway is still possible [34]. Loss of function or gene loss affecting biosynthesis genes or transporters involved in the biosynthesis pathway are rare, but can explain deeper loss of the anthocyanin pigmentation in the Caryophyllales and might be irreversible [35]. Changes in flower color can affect the attraction of pollinating animals and are thus subject to selection. For example, the transition of purple blue to orange-red in *Ipomoea quamoclit* L. probably facilitated the shift from bee pollination to hummingbird pollination [36]. A study in *Silene littorea* discovered that a lack of anthocyanins exclusively in flower petals appeared as a stable polymorphism, but complete loss of anthocyanins in all plant parts was rare suggesting anthocyanins are the target of selection in photosynthetic tissues [37].

The flower pigmentation of *D. purpurea* is not uniform, but characterized by intensely pigmented dark spots on the bottom petals (**Fig. 1a**). Petal spots have been reported in other plant species including *Gorteria diffusa* [38], *Kohleria warszewiczii* [39], *Clarkia gracilis* [40], *Lilium leichtlinii* [41], and *Mimulus guttatus* [42]. Pigmentation spots on petals can be caused by the accumulation of high levels of anthocyanins (cyanidin derivatives) [40, 43]. These spots can be surrounded by a white region characterized by high levels of flavonols, the products of a

pathway competing with anthocyanin biosynthesis [44]. Formation of intensely pigmented spots can be explained by a system involving a local R2R3-MYB activator (NEGAN) and a lateral R3-MYB repressor (RTO) [42]. NEGAN interacts with a bHLH and a WD40 partner (MBW complex) to activate *RTO* expression and RTO represses *NEGAN* expression by sequestering bHLH proteins required for the MBW complex [42]. RTO proteins appear to travel through the endosomal system to other cells, where they act as repressors [42]. Another study in *C. gracilis* reported an anthocyanin activating MYB (CgsMYB12) as the causal gene for the formation of a single large spot [45] indicating that alternative mechanisms have evolved to control the formation of pigmentation spots. Numerous functions have been described for such spots including an attraction and guidance of pollinating insects to increase the number of visits and to make each event more successful [39, 46–49].

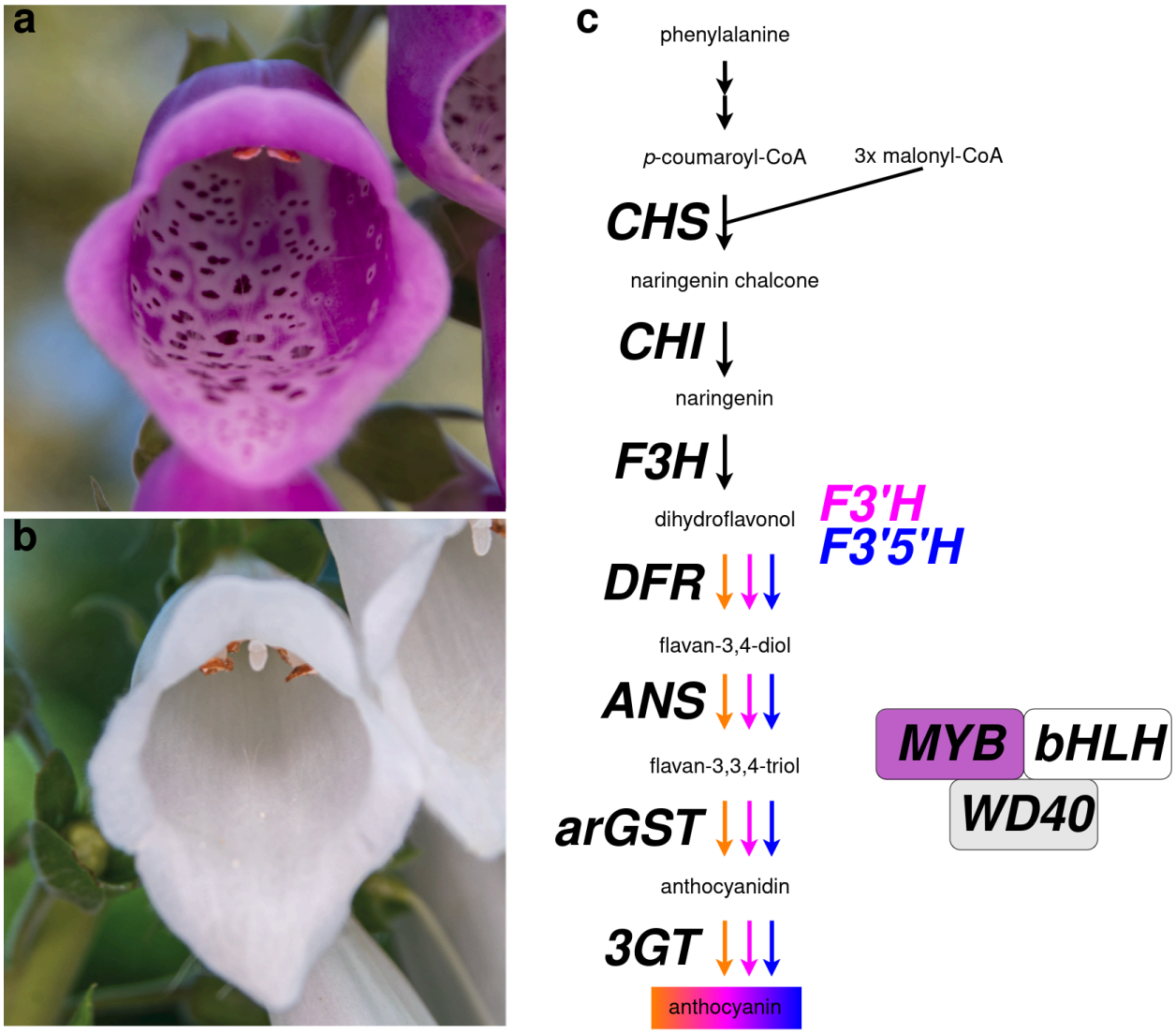

Fig. 1: Pictures showing magenta (a) and white (b) flowers of *Digitalis purpurea* (foxglove). The anthocyanin biosynthesis is displayed as the central branch of the flavonoid biosynthesis (c).

Abbreviations: CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-hydroxylase; DFR, dihydroflavonol 4-reductase; ANS, anthocyanidin synthase; arGST, anthocyanin-related glutathione S-transferase; 3GT, UDP-glucose anthocyanidin 3-O-glucosyltransferase; MYB, Myeloblastosis protein; bHLH, basic helix-loop-helix; WD40, proposed scaffolding protein.

The inflorescence of *D. purpurea* is usually constructed in the form of an indeterminate spike with zygomorphic tubular flowers on the side [50]. But in rather rare cases individuals develop large terminal flowers at the top of the spike, rendering the inflorescence determinate. These phenotypes show two types of flowers in the same plant: the normal zygomorphic flowers at the lower part of the spike and one actinomorphic structured flower at the top (**Fig. 2**). Keeble *et al*. investigated the heredity of this special trait and found it to be inherited recessively [51]. The characterization of the genetics behind the development of a terminal flower in *D. purpurea* has not been investigated yet. The occurrence of terminal flowers in plants with otherwise indeterminate inflorescence architecture has been described in different species like *Antirrhinum* [52] and Arabidopsis [53]. In *Antirrhinum*, a loss of function mutation in the *CEN* (*CENTRORADIALIS*) gene leads to the development of a terminal flower [52]. The functional *CEN* gene is expressed in the subapical meristem and disrupts the initiation of flower development [52]. Therefore, CEN is a flowering repressor. The *Arabidopsis* ortholog *TERMINAL FLOWERING 1* (*TFL1*) plays a similar role [54]. The mutant *tfl1* no longer represses flowering in the apical meristem, so it changes its fate and develops into flower primordia and subsequently into a flower [53, 55]. Both genes encode a protein belonging to the family of phosphatidylethanolamine binding proteins. Therefore, they are not conventional regulators of the transcription or translation of any specific target genes [56]. They have been shown to interact directly with other proteins so they might be part of a more elaborate signaling system [56].

Fig. 2: Three terminal flowers. (A): A magenta terminal flower with purple spots originated from the Botanical Garden of TU Braunschweig. (B and C): White terminal flowering plants.

Here, we report about the genome sequence of *D. purpurea* and the corresponding annotation. As a proof of concept, we demonstrate the power of this genome sequence for the discovery of specialized metabolism pathways by investigating the flavonoid biosynthesis. A comparison of white and magenta flowering individuals revealed molecular differences in the flavonoid biosynthesis genes that could explain the pigmentation loss in the white flowering plants. A candidate variant providing a possible explanation for large terminal flowers and indeterminate flowering architecture was discovered through comparative genomics based on this genome sequence.

Materials & Methods

Plant material and DNA extraction

Digitalis purpurea is a native species in Germany that is widespread at the borders of forests and in gardens. We collected seeds of a white and magenta flowering plant which were already in cultivation in a garden close to these coordinates 51.92351 N, 8.84441 E. No specific permissions were required. Formal identification of the plants was performed by Jakob Horz. Seeds of the investigated plants are now available in the Botanic Gardens of the University of Bonn (BONN-49855).
Plants were cultivated under long day conditions (16h light, 8h darkness) at 20°C. LEDs (Niello®, 300W) were used as the sole light source. Plants were incubated in the dark for two days immediately prior to the sampling of leaves for the DNA extraction to reduce polysaccharide content. Young leaves were harvested for the DNA extraction. The DNA extraction was performed with a CTAB-based protocol as previously described [57]. Quality and purity of the extracted DNA were assessed through NanoDrop measurement, agarose gel electrophoresis, and Qubit measurement. Short fragments (<25kb) were depleted with the Short Read Eliminator kit (Pacific Biosciences) (see [58] for a detailed workflow).

Nanopore sequencing

The library preparation for sequencing was started with 1µg of high molecular weight DNA. The initial DNA repair was performed with two enzyme mixes of the NEBNext® Companion Module Oxford Nanopore Technologies® (ONT) Ligation Sequencing (E7180L) following the suppliers' instructions. All following steps were performed according to the SQK-LSK109 protocol (ONT). Sequencing was performed with MinIONs using R9.4.1 flow cells. Upon blockage of a substantial number of nanopores, a wash step was performed with the EXP-WSH004 following ONT's instructions to recover the blocked nanopores. Afterwards, a fresh library was loaded onto the flow cell. High accuracy basecalling was performed with guppy v6.4.6+ae70e8f (ONT) and minimap2 v2.24-r1122 [59] on a graphical processor unit (GPU) in the de.NBI cloud.

De novo genome sequence assembly

A *de novo* genome sequence assembly was generated based on all long reads of a magenta flowering plant. NextDenovo v2.5.0 [60], Shasta 0.10.0 [61], and Flye 2.9.1-b1780 [62] were deployed to generate assemblies. Based on completeness, continuity, and assumed correctness, an assembly produced by NextDenovo v2.5.0 with read_cutoff = 1k, genome_size = 1g, and seed_depth = 30 was selected as the best representation of the genome. Additional polishing of this assembly was performed with NextPolish v1.4.1 [63] based on read mappings that were generated with minimap2 [59] and samtools v1.15.1 [64]. The genomic BUSCO assessment command contained the arguments '--long --limit 10 -sp arabidopsis', 10^{-3} as e-value cutoff for BLAST, and eudicots_odb10 as the reference [65, 66]. AUGUSTUS v3.2.2 [67, 68] was used for the gene prediction as part of BUSCO. The assembly quality was further evaluated using k-mer analysis YAK (v0.1-r69-dirty) [69] (-k21) and Merqury v1.3 [70] with Meryl v1.4.1 (k=21) based on the ONT reads.

Structural annotation

Paired-end RNA-seq reads (Additional file 1) were retrieved from the Sequence Read Archive via fastq-dump [71] and aligned to the assembled *D. purpurea* genome sequence with STAR v2.5.1b [72, 73] in 2-pass-mode using a minimal similarity of 95% and a minimal alignment length of 90% as previously described [74] to provide hints for the gene prediction process. Gene models were predicted with GeMoMa v1.9 [75, 76] using the following arguments: pc=true pgr=true p=true o=true GeMoMa.c=0.4 GeMoMa.Score=ReAlign Extractor.r=true GAF.f="start=='M' and stop=='*' and (isNaN(score) or score/aa>='0.75')". In addition to RNA-seq hints, the annotation of the following datasets was supplied as hints: *Salvia hispanica* (GCF_023119035.1), *Rehmannia glutinosa* (GCA_016081115.2), *Sesamum indicum* (GCF_000512975.1), *Salvia miltiorrhiza* (GCF_028751815.1), *Penstemon davidsonii* (GCA_034814905.1), and *Penstemon smallii* (GCA_046254845.1). Predicted gene models were filtered via GeMoMa (f="start=='M' and stop=='*' and (isNaN(tie) or tie>0) and tpc>0 and aa>50" atf="tie>0 and tpc>0") and renamed. BUSCO v3.0.2 [65, 77] was used to assess the completeness of the genome sequence and the completeness of the gene prediction. The completeness of predicted gene models was assessed with BUSCO v5.7.1 by running in proteome mode with default parameters.

Functional annotation

Homologs in *Arabidopsis thaliana* were identified for the predicted peptide sequences to transfer annotation terms from this model organism to the *D. purpurea* sequences. Previously developed Python scripts [74, 78] were modified to enable the efficient identification of Reciprocal Best BLAST hits (RBHs) supplemented with best BLAST hits for query sequences that are not

assigned to a RBH partner. BLASTp v2.13.0+ [79, 80] was run with an e-value cutoff of 0.0001. This approach allows an efficient identification of putative orthologs and forms the basis for the transfer of functional annotation details across species borders. *Arabidopsis thaliana* TAIR10/Araport11 annotation terms were automatically transferred to the predicted *D. purpurea* sequences. This workflow is implemented in the Python script construct_anno.py [81]. Enzyme and transporter encoding genes of the flavonoid biosynthesis were identified using KIPEs v3.2.6 with the flavonoid biosynthesis data set v3.4 [6, 82] with additional parameters listed in Additional file 2. Flavonoid biosynthesis controlling MYB transcription factors were annotated with the MYB_annotator v1.0.3 [83]. The following parameters were used for the MYB annotation: initial candidate identification via BLASTp v2.13.0+ [79, 80], alignment construction with MAFFT v7.475 [84], phylogenetic tree construction with FastTree v2.1.10 [85], minimal BLASTp alignment length of 75 amino acids, and 10 neighbors were used for the ingroup/outgroup classification (see Additional file 3 for details). The bHLH transcription factors were annotated using the bHLH_annotator v1.04 [86]. The following parameters were used for the bHLH annotation: identification of initial candidates via BLASTp v2.13.0+ [79, 80], alignment with muscle v5.1 [87], phylogenetic tree construction with FastTree v2 [85], a minimal BLASTp hit similarity of 40%, minimal BLASTp hit alignment length of 80 amino acids, and considering 10 neighbors for the ingroup/outgroup classification (see Additional file 4 for details).

Annotation of transposable elements and repeats

Repetitive sequences were annotated with RepeatModeller2 [88] and RepeatMasker [89]. A species-specific repeat library was generated with RepeatModeller v2.0.7 using the assembled genome sequence as input. Genome-wide repeat masking and content assessment was then performed with RepeatMasker v4.2.2. with the parameters -e rmblast and -lib with the *Digitalis*-specific repeat library. EDTA v2.2.2 [90] was utilized for the genome wide identification and annotation of transposable elements. The parameters for the annotation process were chosen as follows: --species others --step all --overwrite 1 --sensitive 1 --anno 1 --threads 10 --force 1. For evaluating the annotation consistency the parameters were adjusted to: --species others --step anno --overwrite 0 --anno 1 --threads 10 --evaluation 1. For the annotation of TE diagnostic motifs, located on the inserted sequences in *ANS* and TFL1/*CEN*, first the sequence was translated based on putative open reading frames (ORFs) using getorf (EMBOSS v6.6.0.0) [91] with the parameter -minsize 150 to filter out short ORFs. Next the collection of putative polypeptide sequences was subjected to hmmscan (HMMER v3.4) [92] against the Pfam database [93]. The parameters used for hmmscan were --cut_ga and --noali. To find LTRs in the sequence inserted into the *ANS* locus, LTR_Finder v1.07 [94] was applied with default parameters. Two LTR sequences were extracted using bedtools getfasta v.2.31.1 [95] and subsequently aligned with MAFFT v7.525 [84]. The distance between the two sequences was calculated using distmat (EMBOSS v6.6.0.0) [91] with the parameter --nucmethod 2 in order to use Kimura's 2 parameter model [96]. The evolutionary divergence and therefore insertion time of the TE can be estimated with the formula:

$$T = \frac{K}{2*r}$$

With T: Time, K: Substitutions per site and r: Neutral substitution rate.

As neutral substitution rates vary a lot and no values for closely related species of *Digitalis purpurea* are available, the calculation was performed with $r = 1.5 \times 10^{-8}$ a frequently used reported synonymous mutation rate in *Chalcone Synthase* and *Alcohol Dehydrogenase* loci of *A. thaliana* [97–99]. This approach is expected to return a rough estimation, while it is important to note that substitution rates can vary by over a magnitude [100].
As no LTRs were found in the sequence inserted into the *TFL1/CEN* locus, it was searched for possible poly(A) tracts, diagnostic for LINE and SINE non-LTR retrotransposition as reviewed by Kazuhiko Ohshima [101].

Ks analysis and coverage analysis

To detect whole-genome duplication (WGD) events, a Ks analysis was performed. For each gene, the longest polypeptide sequence was selected. A self-BLAST of these polypeptide sequences was then performed using collect_best_BLAST_hits.py [81] with default parameters. The resulting BLAST table, the corresponding CDS file and a gene position file containing genomic coordinates were used as input for MCScan (via JCVI utility library 1.5.11, jcvi.compara.catalog) [102] to identify paralogous gene pairs in a synteny-aware manner. Codon-aware alignments of these gene pairs and calculation of Ks values were subsequently performed with ParaAT 2.0 [103] using the parameters -msa mafft, -format axt, -nogap and -kaks. The resulting ks distribution was plotted with a custom Python script and a Gaussian Mixture Model (GMM) with 5 components was fitted to the ks distribution using the GaussianMixture function from scitkit-learn [104].

To compare the coverage distribution of duplicated and single-copy BUSCO genes, the corresponding gene IDs were separated into two files and their genomic coordinates were extracted from the GFF annotation using the custom Python script extract_coordinates_from_gff.py [105]. Mean per-gene coverage was calculated using bedtools coverage v2.31.1 [95] with the -mean option. Coverage distributions were visualized as violin plots with box overlays using plot_coverage_busco2.py [105]. Peak detection was performed with SciPy's gaussian_kde and find_peaks functions [106]. Statistical differences between the two distributions were assessed with a Mann-Whitney U test.

To assess the genomic heterozygosity, variants were called using bcftools v1.9 [107]. Raw alignment data were processed with bcftools mpileup using the parameters -q 20 (minimum mapping quality), -Q 15 (minimum base quality) and -a FORMAT/DP,FORMAT/AD to annotate read depth and allele depth. The resulting output was piped to bcftools call with the parameters --ploidy 2, -m (multiallelic caller) and -v (output variant sites only) to generate variant calls. High confidence heterozygous single nucleotide polymorphisms (SNPs) were extracted from the vcf file using bcftools view with the filters -i ‘GT=”het”’ and -v snps. The total number of heterozygous SNPs was counted to obtain the numerator for heterozygosity estimation. The number of callable genomic positions was determined using samtools depth with the same quality thresholds applied during variant calling (-q 20 and -Q15). Positions meeting the defined coverage depth criteria were counted to represent callable genome size. Relative

heterozygosity was calculated as the ratio of heterozygous SNPs to the total number of callable genomic positions.

Analysis of gene expression

Fastq-dump [71] was applied to retrieve the FASTQ files of *D. purpurea* RNA-seq experiments from the Sequence Read Archive (SRA). Gene expression was quantified using kallisto v0.44.0 [108] to process all RNA-seq data sets (Additional file 1) based on the predicted coding sequences. Individual count tables were merged with a customized Python script and served as the basis for a visualization of gene expression as previously described [17, 81].

Identification of sequence variants associated with anthocyanin loss

Reads of two magenta flowering plants and two white flowering plants (Additional file 5) were aligned against the reference genome sequence of a magenta flowering plant with minimap v2.24-r1122 [59]. The resulting BAM files were indexed with samtools v1.13 [64]. Candidate genes associated with the flavonoid biosynthesis were manually inspected for systematic sequence differences via Integrative Genomics Viewer (IGV) v2.15.4 [109]. The same method was applied to one terminally flowering individual and three indeterminate flowering individuals.

Identification of paralogous regions and functional assessment of the candidate *ANS* locus

To identify paralogous regions of the genome sequence in comparison to the contig harbouring *ANS* an alignment of ctg00580 against the genome assembly of *Digitalis purpurea* was performed using LAST v.1650 [110] with default parameters. The resulting mapping file was converted from MAF to TSV format using the maf-convert tab and filtered for alignments ≥ 2000 bp with percent identity ≥ 85%. Self-synteny analysis of the *D. purpurea* genome sequence was performed using JCVI v1.5.11. Synteny anchors were identified using jcvi.compara.catalog ortholog with --cscore=0.99 and --no_strip_names, and a microsynteny analysis for the *ANS* locus was conducted using jcvi.compara.synteny mcscan [102].

Genotyping of a *Digitalis purpurea* population

A population of 89 plants was subjected to genotyping via PCR for the discovered mutations putatively associated with the white flower color or the changed flower morphology, respectively. The plants were first grown under long day conditions (16h light, 8h darkness) at 20°C. LEDs (Niello®, 300W) were used as the sole light source during that time. After four months the plants were moved outside the plant cultivation chamber and placed into two garden beds.

DNA for subsequent genotyping via PCR was extracted based on the GABI-Kat CTAB DNA preparation protocol [111] performed with one leaf from a single *D. purpurea* plant per sample. After extraction the DNA samples were stored at -20 °C.

PCR was conducted with oligonucleotides flanking the *TFL1*/*CEN* locus and a 106 bp deletion linked to the mutant *ans* allele at a distance of 1653 bp (Additional file 6, Additional file 7). Sizes of the PCR products were analyzed via agarose gel electrophoresis and revealed the genotypes.

Initial phenotypes of the plants resulting from the *ANS* mutation were derived from the visual inspection of the leaves as *D. purpurea* with a functional *ANS* would show magenta colored petioles (Additional file 8). These phenotypes were later validated by adding information about the flower color.

Results

Genome sequence assembly and structural annotation

A magenta flowering plant (DR1) was selected for sequencing and construction of a reference genome sequence [112]. The rationale behind this selection was the assumption that a magenta flowering plant should have all genes required for the biosynthesis of anthocyanins which are most likely responsible for the magenta flower pigmentation. For comparison, another magenta flowering plant (DR2) and two white flowering plants (DW1, DW2) were subjected to nanopore long read sequencing (Additional file 5).

Different assemblers were evaluated with respect to their performance in this study (**Table 1**). The assembly produced by NextDenovo2 was selected as the representative *D. purpurea* genome sequence due to the large assembly size of 940 Mbp, the high continuity (N50=4.3Mbp), and the high completeness of 96% of complete BUSCOs. Both quality values for the NextDenovo2 assembly estimated with YAK and Merqury resulted in 27.7 corresponding to ~99.8 % base level accuracy.

Table 1: Comparison of assembler performance for the analysis of *Digitalis purpurea* long read sequencing data of the magenta flowering plant DR1.

Assembly ID	**DR1_v1**	**DR1_v2**	**DR1_v3**
Assembler	NextDenovo (v2.5.0)	Shasta (v0.10.0)	Flye (2.9.1-b1780)
Size	940 Mbp	773 Mbp	1379 Mbp
# contigs	684	4883	3738
N50	4.3 Mbp	0.9 Mbp	1.9 Mbp

N90	464 kbp	71 kbp	132 kbp
BUSCO	C:95.7% [S:63.5%,D:32.2%], F:0.6%, M:3.7% n:2326	C:94.4% [S:68.7%,D:25.7%], F:0.6%, M:5.0%, n:2326	C:25.9% [S:23.3%,D:2.6%], F:6.0%, M:68.1% n:2326

Only the representative genome sequence DR1_v1 was subjected to a gene prediction process that resulted in 35,916 protein encoding genes [113]. A completeness check revealed 96.3% complete BUSCO genes, of which a high proportion (55.8 %) were classified as duplicated. Comparison of mean read depth between single-copy and duplicated BUSCO genes revealed clearly distinct coverage distributions (Additional file 9 (Fig. S1)). Single-copy BUSCOs exhibited a single dominant peak at 31.8 x coverage, closely matching the average genome-wide coverage (33.6 x) (Additional file 9). In contrast, duplicated BUSCOs showed a bimodal coverage distribution, with one peak at 30.4 x and a second peak 14.8 x. The higher-coverage peak represents true duplicated loci (at merged haplophases), whereas the lower-coverage peak represents resolved haplophases. Together, these results indicate that the high proportion of duplicated BUSCO genes (55.8 %) can be explained by a combination of a recent WGD event and a partial haplotype separation in the assembly.

Ks analysis of syntenic paralog pairs identified a major peak at Ks = 0.21 and a less pronounced shoulder at Ks = 0.34, as inferred from Gaussian mixture modelling (**Fig. 3** and Additional File 9 (Fig. S2, Fig. S3)). Components at substantially higher Ks values were not considered further, as they likely reflect background noise rather than large scale duplication events [114]. Based on a Digitalis-specific repeat library, RepeatMasker estimated that interspersed repetitive sequences comprise 73.8 % of the genome sequence (Additional File 11). Annotation of transposable elements and repeats using EDTA revealed that TE derived sequences account for 67.4 % of the genome sequence. The majority of repeats are LTR retrotransposons which alone comprise 62.4 % of the assembly (Additional File 12). Genome-wide heterozygosity was estimated at 0.7 %, based on 2,028,683 heterozygous SNPs across 304,433,041 bp of callable sites in the genome sequence.

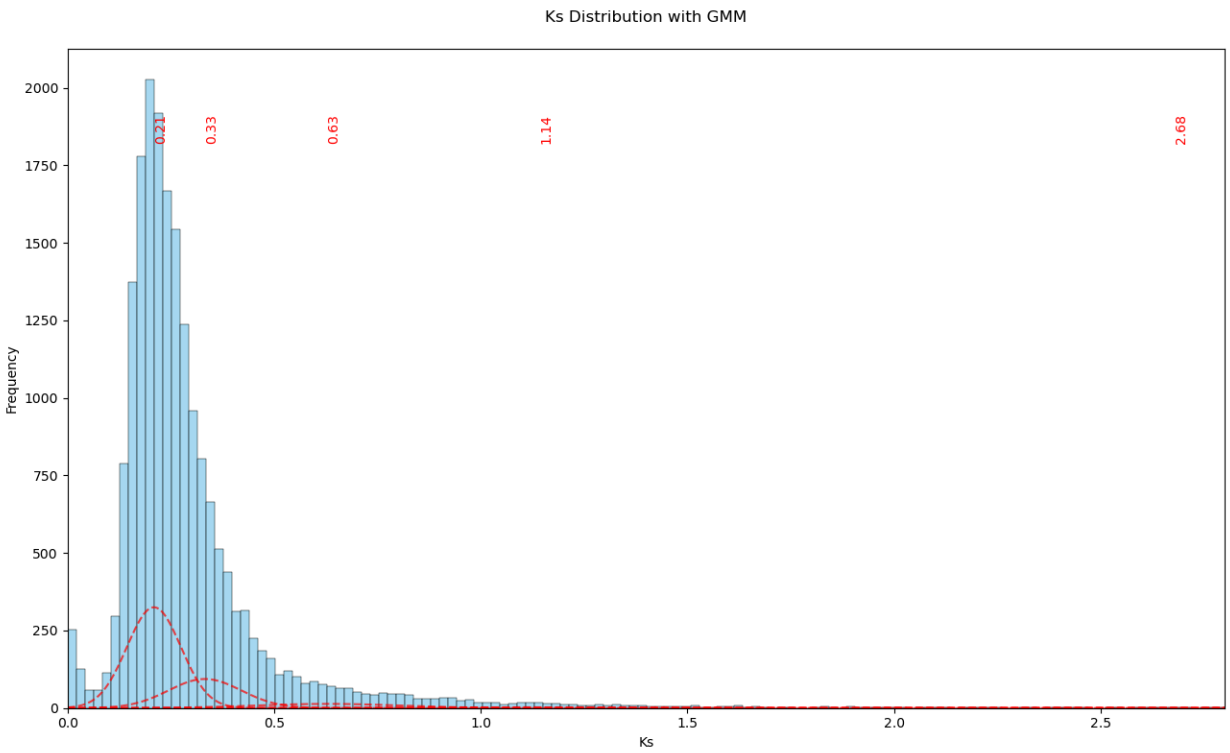

Fig. 3: Ks distribution across syntenic gene pairs. Red dashed curves represent the individual components of a Gaussian mixture model (5 components) fitted to the Ks distribution, with red labels indicating the inferred Ks peak positions. The x-axis is limited to Ks ≤ 2.8 to improve visualization.

Anthocyanin biosynthesis genes in *Digitalis purpurea*

Anthocyanins are well known pigments responsible for the coloration of many plant structures - especially flowers. Therefore, the genes encoding enzymes of the anthocyanin biosynthesis (*CHS*, *CHI*, *F3H*, *F3'H*, *DFR*, *ANS*, *arGST*, *3GT*) were identified (**Fig. 4**; Additional file 13). Additionally, the transcription factors controlling the anthocyanin biosynthesis genes in *D. purpurea* were investigated. Genes belonging to the MYB75/PAP1 (DP112203, DP103545, DP109418) and MYB123/TT2 (DP102989) lineages were identified. In addition, a candidate for the bHLH partner in the anthocyanin biosynthesis MBW regulatory complex (bHLH42/TT8) was detected in *D. purpurea* (DP105472). An ortholog of the WD40 gene *TTG1* (DP111282) was also identified in the phylogenetic analysis. In summary, all genes expected for a complete anthocyanin biosynthesis pathway were successfully identified in the reference genome sequence DR1_v1.

Additional support for candidate genes comes from a gene expression analysis that was conducted across a wide range of different samples (**Fig. 4**, Additional file 1, Additional File 14). Anthocyanin biosynthesis genes are expected to be highly expressed in pigmented structures like flowers and in particular within the intensely pigmented purple spots on the lower flower

petal. A consideration of the gene expression pattern enabled a reduction of the candidate gene set to those gene copies which appear active in the anthocyanin-pigmented plant organs.

Genes encoding enzymes for all steps of the flavonoid and anthocyanin biosynthetic pathways were found to contain at least one paralog that is transcriptionally active in petal tissues. Two paralogous *F3'H* genes exhibited expression above baseline levels. *F3'Ha* (DP114378) was strongly active in floral tissues, whereas *F3'Hb* (DP125257) showed elevated expression primarily in root tissues (Additional File 14). Although roots typically do not accumulate anthocyanins, *F3'Hb* is reported here for completeness.

The putative anthocyanin-related transcription factor MYB75 is represented by three paralogous genes that displayed distinct expression profiles. *DpMYB75a* was active in the lower, upper, and side petal tissues but not in the spot sample. In contrast, *DpMYB75b* was strongly active in spot tissues and, consequently, also showed elevated expression in the lower petal samples, which include the spot region. *DpMYB75c* exhibited a comparatively uniform expression pattern across young floral tissues, with a modest increase in the spot samples.

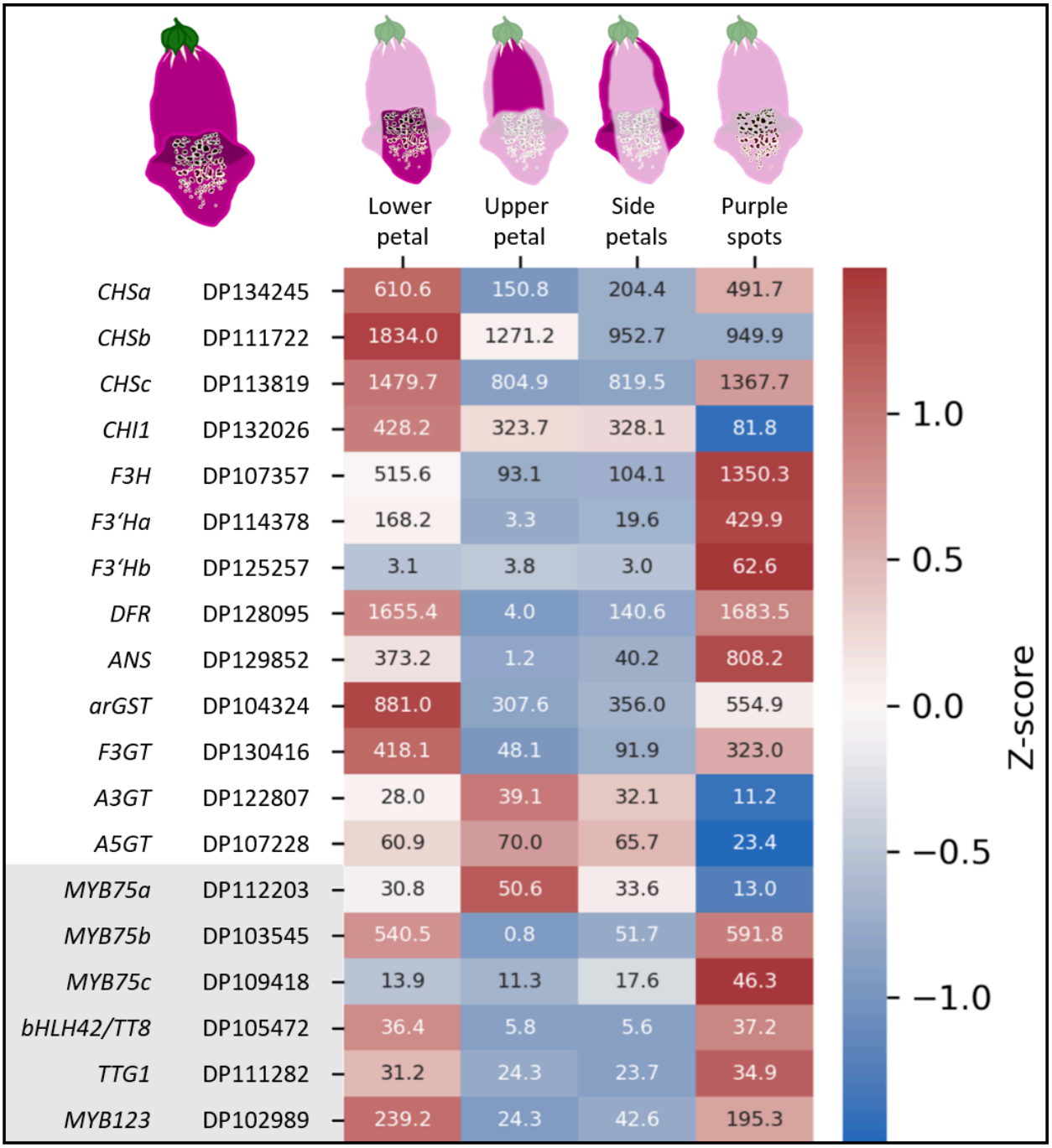

Fig. 4: Expression plot showing the activity of anthocyanin biosynthesis associated genes across different petal tissues. The coloration indicates the z-score, the numbers inside the squares are the corresponding TPM-values. Abbreviations: *CHS*, chalcone synthase; *CHI*, chalcone isomerase; *F3H*, flavanone 3-hydroxylase; *F3'H*, flavonoid 3'-hydroxylase; *DFR*, dihydroflavonol 4-reductase; *ANS*, anthocyanidin synthase; *arGST*, anthocyanin-related glutathione S-transferase; *3GT*, anthocyanidin 3-O-glucosyltransferase; *F3GT*, UDP-glucose flavonoid 3-O-glucosyltransferase; *A3GT*, UDP-glucose anthocyanidin 3-O-glucosyltransferase; *A5GT*, UDP-glucose anthocyanidin 5-O-glucosyltransferase; MYB, Myeloblastosis; bHLH, basic helix-loop-helix; *TTG1*, *TRANSPARENT TESTA GLABRA 1*.

Sequence variants underlying pigmentation differences

All identified candidate genes associated with the anthocyanin biosynthesis were manually inspected in a read mapping comparing magenta flowering and white flowering plants. The most striking difference is an insertion of approximately 14.2 kb in the *ANS* gene (**Fig. 5**). This insertion is homozygous in the two white flowering plants, while the magenta flowering plants harbor one *ANS* allele with and one without this insertion and are therefore heterozygous (Additional file 15 (Table S3)). Although the reference genome sequence was derived from a magenta flowering plant, read mapping indicates that only the mutant allele is represented in the reference assembly, due to the heterozygous genotype of the individual. Since the reference genome sequence DR1_v1 contains the 14.2 kb sequence, the automatic prediction of the *ANS* gene model is inaccurate. The inserted sequence was misannotated as an intron of the *ANS* gene because no RNA-seq reads were aligned to the inserted region. However, in the wildtype allele, the first exon extends across the position of the insertion, indicating that the sequence cannot be removed by splicing in the mutant allele. This strongly suggests that the insertion disrupts the normal *ANS* transcript structure rather than representing an intronic region (**Fig. 5**). The inserted sequence was not annotated as a TE by EDTA. However, five Pfam domains located on five individual ORFs were found by hmmscan: zf-CCHC (Zinc knuckle), Pol_BBD (Pol polyprotein, beta-barrel domain), gag_pre_integrs (GAG-pre-integrase domain), SH3_retrovirus (Retroviral polymerase SH3-like domain) and RVT_2 (Reverse transcriptase) (Additional File 16). ORFs encoding RNase H domain and protease domains, essential for retrotransposons, were not found. Dating of the LTR sequences revealed a distance of 0.0265 substitutions per site. The inference of absolute insertion times with absent lineage-specific neutral substitution rates is imprecise. Nevertheless, LTR divergence has been widely used to infer relative TE age. With an assumed neutral substitution rate of 1.5×10^{-8} mutations per site per year, the insertion can be estimated to have taken place roughly 8.8×10^{5} years ago. In the genomic region containing the transposable element (TE) insertion, the coverage plots of magenta-flowering plants do not display exactly half the read depth of adjacent genomic regions, as would be expected for heterozygous individuals (**Fig. 5**). Instead, the coverage is slightly elevated. Similarly, in white-flowering plants, the coverage across the insertion site is also somewhat higher than that of the surrounding sequences. These deviations from the expected pattern are likely caused by misaligned reads originating from additional copies of the TE elsewhere in the genome.

To analyse the presence of additional *ANS* copies or homologous genomic regions resulting from the genome duplication, the contig harbouring *ANS* was aligned against the whole genome sequence. However, no alternative hits were identified (Additional file 17). While a microsynteny analysis revealed a syntenic region of the genes surrounding *ANS* on contig ctg000280, no *ANS* gene was identified on this contig (Additional file 18). This lack of detection in a highly continuous and complete genome sequence indicates that no alternative copy of *ANS* is present within the *D. purpurea* genome.

Genotyping of 89 plants with respect to the *ANS* locus identified 14 plants with the homozygous wildtype allele, 31 plants with a heterozygous genotype, and 44 plants with the homozygous mutant allele. In the genotyped population, 61 plants have flowered. The flowering phenotype was found to fit the genotype in 57 of these 61 plants, making up 93.4 %. Four plants showed a phenotype mismatching the genotyping results. Two plants flowered white while being heterozygous for the mutation and two plants showed magenta flowers but had a homozygous mutant genotype (Additional File 15 (Table S1)).

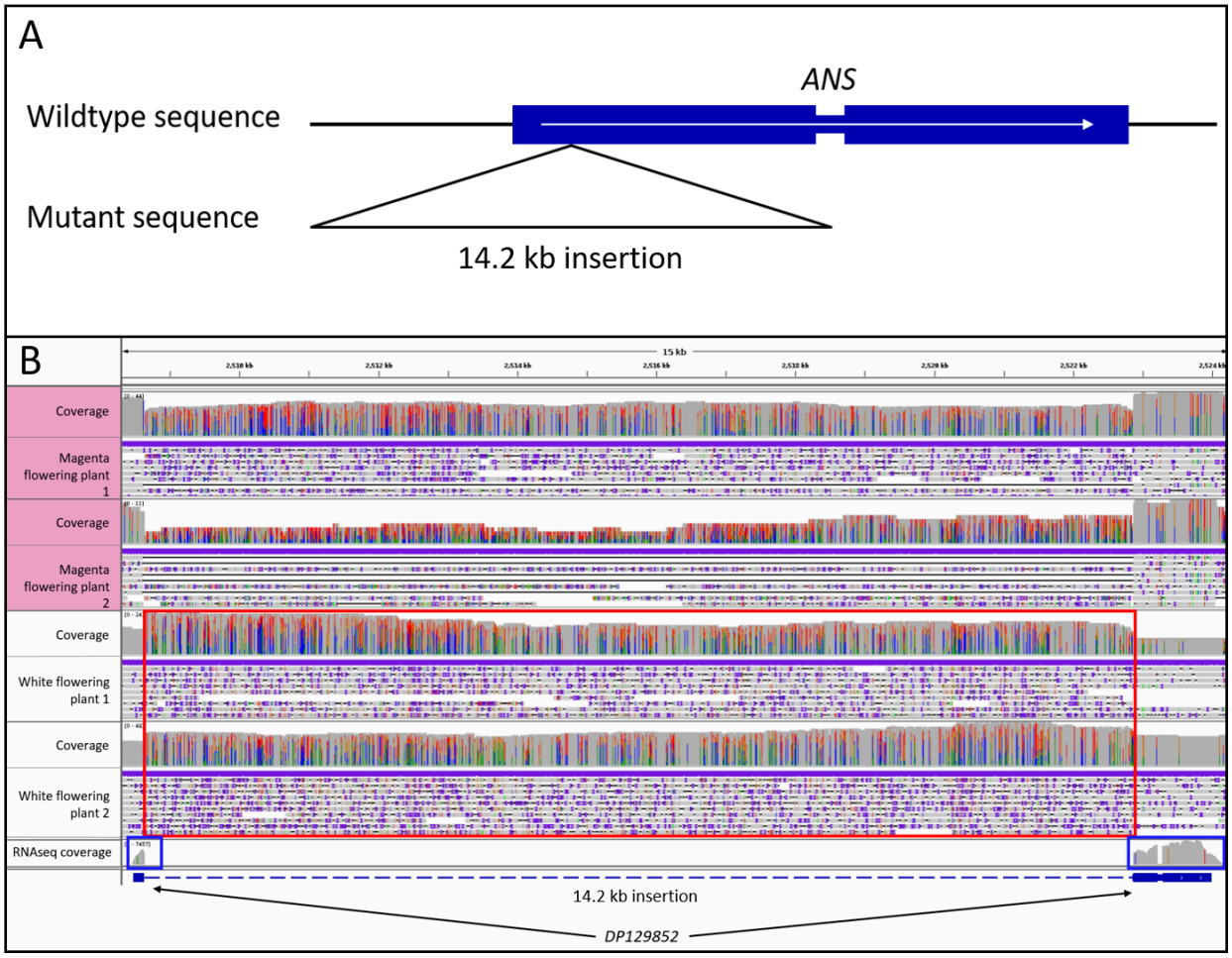

Fig. 5: Comparison of the *ANS* locus in magenta and white flowering *D. purpurea* plants revealed a large insertion that appears homozygous in the white flowering plants. The insertion is expected to be a loss-of-function mutation that results in the loss of ANS enzymatic activity and thus a block in the anthocyanin biosynthesis ultimately resulting in the white flower phenotype. **A:** Schematic of the insertion. **B:** Screenshot from IGV with ONT read mappings of two magenta flowering plants and two white flowering ones at the *ANS* locus. The red box highlights the inserted reads. The blue boxes highlight RNA-seq read mapping coverage to corroborate the gene annotation.

Sequence variants leading to the formation of a large terminal flower

In the search for the genetic basis of the formation of large terminal flowers, the sequencing reads of one plant with a terminal flower were compared to three plants showing wildtype flower morphology. Upon investigation of the *Digitalis* orthologs of the *Arabidopsis TFL1* a 14.3 kb insertion was found starting only 4 base pairs downstream of the start codon of the gene and leading to a premature stop codon 4 codons further downstream (**Fig. 6**). The genotypes of the sequenced individuals at the *TFL1/CEN* locus can be found in Additional file 15 (Table S3). No LTR-sequences were found in the reassembled insert sequence. However, two significant hits for both the integrase domain (PF00665) and the Reverse Transcriptase domain (PF07727) were discovered by hmmscan. Also, a region of 131 bp containing two poly(A) sequences (50 and 48 bp in length, respectively) separated by 33 bp was found.

Out of 24 characterized plants, 23 showed the phenotype fitting the predicted genotype. While 17 plants showed the wildtype floral architecture and contained at least one wildtype allele, 7 plants developed terminal flowers. Of those plants with large terminal flowers, six had the homozygous mutant genotype and one appeared to have a heterozygous genotype (Additional File 15 (Table S2)).

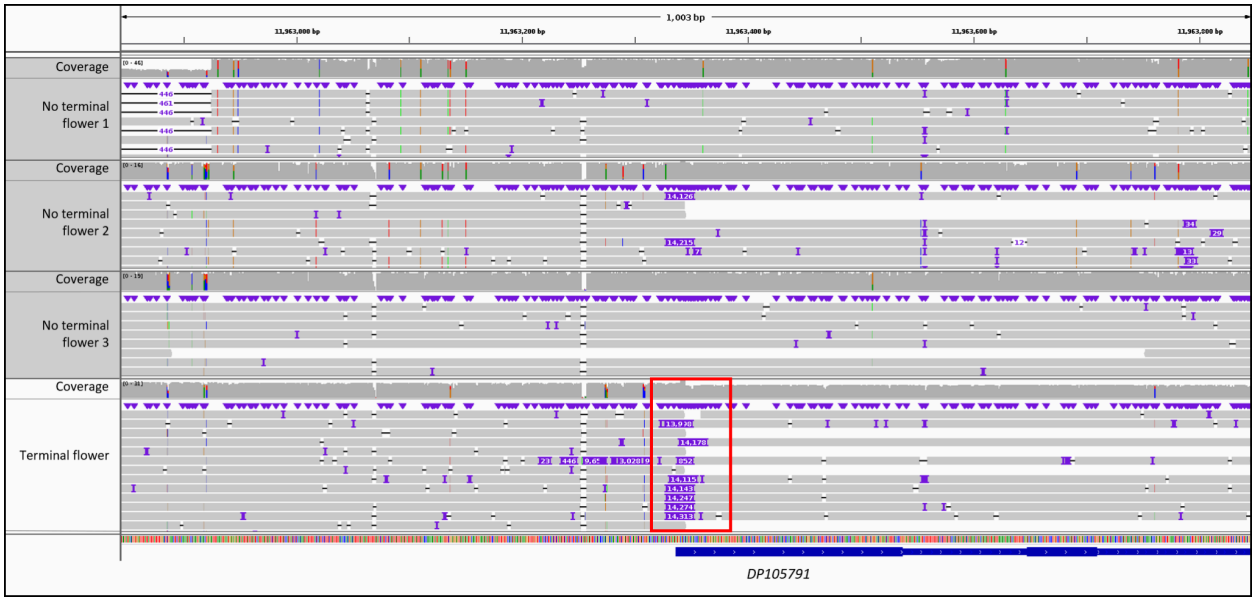

Fig. 6. Genomic region showing the insertion of 14 kb in the *DpFL1/CEN* (DP105791) in the reads of the plant with a terminal flower (red box). The plants with indeterminate flowering architecture show at least some reads without the insert.

Discussion

The genome sequence of *D. purpurea* and the corresponding annotation of protein encoding genes enabled the identification of genes associated with the anthocyanin biosynthesis. Long read sequencing of two white flowering individuals enabled the discovery of sequence variants that might explain the lack of pigmentation. *ANS* encodes the anthocyanidin synthase, a crucial enzyme for the anthocyanin biosynthesis, and shows a homozygous insertion of a large DNA fragment in both white flowering plants, while the two magenta flowering plants show at least one allele without this insertion. The consequence of this insertion is most likely a loss of ANS activity that disrupts the anthocyanin biosynthesis. A loss of ANS activity fits all observations regarding the presence/absence of flavonoids in *D. purpurea*.

The insertion was found to contain TE-associated motifs like flanking LTRs as well as motifs of reverse transcriptase domain and GAG domain. Furthermore zinc knuckle, beta-barrel and SH3 retrovirus domains were detected. The lack of RNase H and protease domains combined with the fact that the detected motifs are not encoded by one continuous ORF but by five individual overlapping ORFs indicate that this TE is no longer functional and most likely heavily degenerated. This fits the obtained LTR dating results placing the insertion almost 900,000 years ago. When considering slower substitution rates available for plants, this time could increase drastically to well above 5 Mya [100]. Due to missing linear specific substitution rates the exact time remains unknown. It can be said that the insertion of the TE and thereby the occurrence of white flowering *Digitalis purpurea* plants predates the first horticultural efforts made by humans [115].

We noticed that flavonols are produced in the white flowering plants (Additional file 19). There is an overlap regarding the enzymes required for flavonol and anthocyanin biosynthesis. CHS, CHI, F3H, and F3'H are shared between both pathways. The presence of flavonols indicates that the genes *CHS*, *CHI*, *F3H*, and *F3'H* are functional, because they are required for the flavonol biosynthesis [116–120]. This observation reduces the set of remaining structural genes in the anthocyanin pathways to *DFR*, *ANS*, *arGST*, and *UFGT*. While we cannot completely rule out that *DFR*, *arGST*, or *UFGT* contribute to the loss of anthocyanin pigmentation, we have not observed any evidence for systematic sequence variants in these genes. Several previous studies discovered loss-of-function mutations in *ANS* as the cause of anthocyanin pigmentation loss [121–123]. The above described *ans* mutation could be a possible explanation for the lack of pigments in white *D. purpurea* flowers. Because the tested population did not show a perfect correlation between *ANS* genotype and floral phenotype, it remains possible that additional mechanisms contribute to the absence of pigmentation in white-flowering *D. purpurea* individuals. A further plausible explanation is that two plants grew in close proximity and physically merged, thereby compromising the representativeness of the sampled tissue. As transcription factors are often implicated in the loss of anthocyanin pigmentation [34], the identified candidate genes were checked for sequence variants associated with the anthocyanin loss. However, no variants fitting the phenotypes could be observed in these genes. Nevertheless, three paralogous *DpMYB75* candidates were found to have differential expression patterns in petal tissues. DpMYB75b is strongly active in the dark purple spots which

suggests its involvement in the regulation of spot pigmentation DpMYB75a is expressed in the remaining floral samples but not in the spot sample. This suggests the existence of two separate regulatory complexes involved in the pigmentation of *Digitalis* flowers: One activating spot pigmentation (DpMYB75b) and one activating background pigmentation (DpMYB75a). Whether or not the absence of color and spot pigmentation in white flowering plants has any effect on pollination efficiency remains to be tested. However, there are many examples of plants where different visual cues in the flower lead to increased pollination efficiency [124–126]. The visual cues can consist of flower forms indicating landing sites for insects, simply pigmented flowers letting the flower stand out against the background or contrasting patterns indicating landing-position and -direction for insects on the flower [127, 128]. This suggests that the pigmentation patterns in *D. purpurea* flowers might contribute to increased attractiveness of the flower for pollinators. The contrasting spots on the lower petal possibly indicate a landing zone and the two UV fluorescent stripes at the base of the connation of the side petal with the lower petal may act as nectar guides (Additional file 19).

The present study focuses on explaining the completely white flowering phenotype (**Fig. 1**), but there are also cases of white flowers with magenta spots in *D. purpurea*. Similar patterns have been described before in other plant species [42]. It seems plausible to assume that a similar local activator/lateral repressor system comprising NEGAN and RTO also underlies the pigmentation pattern in *D. purpurea*. Further investigations are required to test this hypothesis.

DpTFL1/CEN is an ortholog of *AmCEN* and *AtTFL1*, genes crucial for the development of an indeterminate inflorescence architecture. Loss-of-function mutations of these genes have been shown to lead to large terminal flowers [52, 54]. The system responsible for inflorescence architecture seems to be highly conserved across plant species [56], suggesting a similar system in *D. purpurea*. Genotyping of a *D. purpurea* population for the *DpTFL1/CEN* locus revealed that plants exhibiting terminal flowers predominantly carry the homozygous mutant genotype (6/7), with one being heterozygous at the *DpTFL1/CEN* locus, suggesting possible mosaicism, misclassification or other influencing factors. If it is indeed only the homozygous mutant genotype that leads to the development of a determinate inflorescence architecture, this would corroborate the findings of Keeble *et al*. [51].
This study has two main limitations: (1) the number of sequenced individuals was limited (n=4), which constrains the ability to capture population-level genetic variation, and (2) PCR-based genotyping showed minor inconsistencies with the respective phenotypes (4 of 61 individuals for the *ANS* and 1 of 24 for the *TFL1/CEN* locus) indicating that the marker does not fully explain phenotypic variation or that other biological or technical factors contribute to mismatches. While the overall consistency was high, these factors should be considered when interpreting these results.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All analyzed RNA-seq data sets are publicly available from the Sequence Read Archive (SRA) (Additional file 1). ONT sequencing data of *Digitalis purpurea* are available from the European Nucleotide Archive (ENA) under PRJEB63449 (Additional file 5).The assembled genome sequence of DR1 and the corresponding annotation are available via bonndata: https://doi.org/10.60507/FK2/4KUKXI [113]. Custom Python scripts developed for this project are available via zenodo: https://doi.org/10.5281/zenodo.18637460.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

BP planned the study. JH, KW, andRF performed the sequencing. JH and BP wrote the software and performed the bioinformatic analysis. JH performed genotyping experiments. JH, KW, RF, and BP interpreted the results, and wrote the manuscript. All authors have read the final version of the manuscript and approved its submission.

Acknowledgements

This work was supported by the de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) and ELIXIR-DE (Forschungszentrum Jülich and W-de.NBI-001, W-de.NBI-004, W-de.NBI-008, W-de.NBI-010, W-de.NBI-013, W-de.NBI-014, W-de.NBI-016, W-de.NBI-022). We also thank all students participating in our 'Data Literacy in Genome Research' course and members of the research group Plant Biotechnology and Bioinformatics for discussion and support. We thank Hanna Marie Schilbert for spotting large terminal flowers on *Digitalis* and contributing to the brainstorming for experiments. We acknowledge support from Project DEAL and the University of Bonn for open access publication.

Additional files

Additional file 1: *Digitalis purpurea* RNA-seq data sets that delivered hints for the gene prediction process.
Additional file 2: Documentation of all parameters and input files used for the annotation of the flavonoid biosynthesis genes with KIPEs3.
Additional file 3: Documentation of all parameters and input files used for the annotation of the MYB transcription factor gene family with the MYB_annotator.
Additional file 4: Documentation of all parameters and input files used for the annotation of the bHLH transcription factor gene family with the bHLH_annotator.
Additional file 5: *Digitalis purpurea* long read sequencing runs with corresponding accessions assigned by the European Nucleotide Archive (ENA).
Additional file 6: Design of the genotyping PCRs for the *ANS* and *TFL1/CEN* locus.
Additional file 7: PCR cycle program for the *DpCEN* and *DpANS* genotyping.
Additional file 8: *D. purpurea* plants showing magenta colored petioles.
Additional file 9: coverage plots of single copy and duplicated BUSCO genes; Ks plot of all syntenic gene pairs and Ks plot of all syntenic gene pairs in y log scale.
Additional file 10: Histogram showing the read depth across the entire assembly. A coverage of 20x suggests a successful separation of haplophases, while a coverage of 40x suggests that haplophases have been merged.
Additional file 11: Summary file output by RepeatMasker.
Additional file 12: Summary file output by EDTA.
Additional file 13: *Digitalis purpurea* candidate genes associated with the anthocyanin biosynthesis.
Additional file 14:Expression plot showing the activity of anthocyanin biosynthesis associated genes across different samples.
Additional file 15: Documentation of the genotyping results of a *D. purpurea* population at the loci *ANS* and *TFL1/CEN* and and the corresponding phenotypes as well as the genotypes of the four sequenced individuals.
Additional file 16: Map of the TE inserted into *ANS* depicting LTR sequences and hmmer motif-hits as well as their corresponding ORFs.

Additional file 17: Single LAST hit produced by searching the Assembly against the contig harboring the *ANS* gene.
Additional file 18: Microsynteny plot showing collinear relationships between the contig harboring the *ANS* gene and a detected syntenic region.
Additional file 19: Magenta flower (left) and white flower (right) under UV-illumination. The two upper petals were removed to open the flower and expose the inside of the remaining flower.

References

1. Kreis W. The Foxgloves (Digitalis) Revisited. Planta Med. 2017;83:962–76. https://doi.org/10.1055/s-0043-111240.

2. Bucca A. Role of Digoxin in Heart Failure. In: Vasan RS, Sawyer DB, editors. Encyclopedia of Cardiovascular Research and Medicine. Oxford: Elsevier; 2018. p. 323–6. https://doi.org/10.1016/B978-0-12-809657-4.99654-6.

3. Carroll E, Ravi Gopal B, Raghavan I, Mukherjee M, Wang ZQ. A cytochrome P450 CYP87A4 imparts sterol side-chain cleavage in digoxin biosynthesis. Nat Commun. 2023;14:4042. https://doi.org/10.1038/s41467-023-39719-4.

4. Winkel-Shirley B. Flavonoid Biosynthesis. A Colorful Model for Genetics, Biochemistry, Cell Biology, and Biotechnology. Plant Physiol. 2001;126:485–93. https://doi.org/10.1104/pp.126.2.485.

5. Pucker B, Selmar D. Biochemistry and Molecular Basis of Intracellular Flavonoid Transport in Plants. Plants. 2022;11:963. https://doi.org/10.3390/plants11070963.

6. Pucker B, Reiher F, Schilbert HM. Automatic Identification of Players in the Flavonoid Biosynthesis with Application on the Biomedicinal Plant Croton tiglium. Plants. 2020;9:1103. https://doi.org/10.3390/plants9091103.

7. Eichenberger M, Schwander T, Hüppi S, Kreuzer J, Mittl PRE, Peccati F, et al. The catalytic role of glutathione transferases in heterologous anthocyanin biosynthesis. Nat Catal. 2023;6:927–38. https://doi.org/10.1038/s41929-023-01018-y.

8. Goodman CD, Casati P, Walbot V. A Multidrug Resistance–Associated Protein Involved in Anthocyanin Transport in Zea mays. Plant Cell. 2004;16:1812–26. https://doi.org/10.1105/tpc.022574.

9. Zhao J, Dixon RA. MATE Transporters Facilitate Vacuolar Uptake of Epicatechin 3′-O-Glucoside for Proanthocyanidin Biosynthesis in Medicago truncatula and Arabidopsis. Plant Cell. 2009;21:2323–40. https://doi.org/10.1105/tpc.109.067819.

10. Mueller LA, Goodman CD, Silady RA, Walbot V. AN9, a Petunia Glutathione S-Transferase Required for Anthocyanin Sequestration, Is a Flavonoid-Binding Protein. Plant Physiol. 2000;123:1561–70. https://doi.org/10.1104/pp.123.4.1561.

11. Nozue M, Yasuda H. Occurrence of anthocyanoplasts in cell suspension cultures of sweet potato. Plant Cell Rep. 1985;4:252–5. https://doi.org/10.1007/BF00269370.

12. Grotewold E, Chamberlin M, Snook M, Siame B, Butler L, Swenson J, et al. Engineering Secondary Metabolism in Maize Cells by Ectopic Expression of Transcription Factors. Plant Cell. 1998;10:721–40. https://doi.org/10.1105/tpc.10.5.721.

13. Zhang H, Wang L, Deroles S, Bennett R, Davies K. New insight into the structures and formation of anthocyanic vacuolar inclusions in flower petals. BMC Plant Biol. 2006;6:29. https://doi.org/10.1186/1471-2229-6-29.

14. Poustka F, Irani NG, Feller A, Lu Y, Pourcel L, Frame K, et al. A Trafficking Pathway for Anthocyanins Overlaps with the Endoplasmic Reticulum-to-Vacuole Protein-Sorting Route in Arabidopsis and Contributes to the Formation of Vacuolar Inclusions. Plant Physiol. 2007;145:1323–35. https://doi.org/10.1104/pp.107.105064.

15. Luo P, Ning G, Wang Z, Shen Y, Jin H, Li P, et al. Disequilibrium of Flavonol Synthase and Dihydroflavonol-4-Reductase Expression Associated Tightly to White vs. Red Color Flower Formation in Plants. Front Plant Sci. 2016;6.

16. Johnson ET, Ryu S, Yi H, Shin B, Cheong H, Choi G. Alteration of a single amino acid changes the substrate specificity of dihydroflavonol 4-reductase. Plant J Cell Mol Biol. 2001;25:325–33. https://doi.org/10.1046/j.1365-313x.2001.00962.x.

17. Choudhary N, Pucker B. Conserved amino acid residues and gene expression patterns associated with the substrate preferences of the competing enzymes FLS and DFR. PLOS ONE. 2024;19:e0305837. https://doi.org/10.1371/journal.pone.0305837.

18. Weisshaar B, Jenkins GI. Phenylpropanoid biosynthesis and its regulation. Curr Opin Plant Biol. 1998;1:251–7. https://doi.org/10.1016/s1369-5266(98)80113-1.

19. Li S. Transcriptional control of flavonoid biosynthesis. Plant Signal Behav. 2014;9:e27522. https://doi.org/10.4161/psb.27522.

20. Xu W, Dubos C, Lepiniec L. Transcriptional control of flavonoid biosynthesis by MYB–bHLH–WDR complexes. Trends Plant Sci. 2015;20:176–85. https://doi.org/10.1016/j.tplants.2014.12.001.

21. Lloyd A, Brockman A, Aguirre L, Campbell A, Bean A, Cantero A, et al. Advances in the MYB–bHLH–WD Repeat (MBW) Pigment Regulatory Model: Addition of a WRKY Factor and Co-option of an Anthocyanin MYB for Betalain Regulation. Plant Cell Physiol. 2017;58:1431–41. https://doi.org/10.1093/pcp/pcx075.

22. Ramsay NA, Glover BJ. MYB-bHLH-WD40 protein complex and the evolution of cellular diversity. Trends Plant Sci. 2005;10:63–70. https://doi.org/10.1016/j.tplants.2004.12.011.

23. Gonzalez A, Zhao M, Leavitt JM, Lloyd AM. Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in Arabidopsis seedlings. Plant J. 2008;53:814–27. https://doi.org/10.1111/j.1365-313X.2007.03373.x.

24. Feller A, Machemer K, Braun EL, Grotewold E. Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. Plant J. 2011;66:94–116. https://doi.org/10.1111/j.1365-313X.2010.04459.x.

25. Albert NW, Butelli E, Moss SMA, Piazza P, Waite CN, Schwinn KE, et al. Discrete bHLH transcription factors play functionally overlapping roles in pigmentation patterning in flowers of Antirrhinum majus. New Phytol. 2021;231:849–63. https://doi.org/10.1111/nph.17142.

26. Airoldi CA, Hearn TJ, Brockington SF, Webb AAR, Glover BJ. TTG1 proteins regulate circadian activity as well as epidermal cell fate and pigmentation. Nat Plants. 2019;5:1145–53. https://doi.org/10.1038/s41477-019-0544-3.

27. Gonzalez A, Brown M, Hatlestad G, Akhavan N, Smith T, Hembd A, et al. TTG2 controls the developmental regulation of seed coat tannins in Arabidopsis by regulating vacuolar transport steps in the proanthocyanidin pathway. Dev Biol. 2016;419:54–63. https://doi.org/10.1016/j.ydbio.2016.03.031.

28. Borevitz JO, Xia Y, Blount J, Dixon RA, Lamb C. Activation Tagging Identifies a Conserved MYB Regulator of Phenylpropanoid Biosynthesis. Plant Cell. 2000;12:2383–93.

29. Tohge T, Nishiyama Y, Hirai MY, Yano M, Nakajima J, Awazuhara M, et al. Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. Plant J Cell Mol Biol. 2005;42:218–35. https://doi.org/10.1111/j.1365-313X.2005.02371.x.

30. Albert NW, Davies KM, Lewis DH, Zhang H, Montefiori M, Brendolise C, et al. A conserved network of transcriptional activators and repressors regulates anthocyanin pigmentation in eudicots. Plant Cell. 2014;26:962–80. https://doi.org/10.1105/tpc.113.122069.

31. Quattrocchio F, Wing J, Woude K van der, Souer E, Vetten N de, Mol J, et al. Molecular Analysis of the anthocyanin2 Gene of Petunia and Its Role in the Evolution of Flower Color. Plant Cell. 1999;11:1433–44. https://doi.org/10.1105/tpc.11.8.1433.

32. Streisfeld MA, Young WN, Sobel JM. Divergent selection drives genetic differentiation in an R2R3-MYB transcription factor that contributes to incipient speciation in Mimulus aurantiacus. PLoS Genet. 2013;9:e1003385. https://doi.org/10.1371/journal.pgen.1003385.

33. Wheeler LC, Walker JF, Ng J, Deanna R, Dunbar-Wallis A, Backes A, et al. Transcription Factors Evolve Faster Than Their Structural Gene Targets in the Flavonoid Pigment Pathway. Mol Biol Evol. 2022;39:msac044. https://doi.org/10.1093/molbev/msac044.

34. Marin-Recinos MF, Pucker B. Genetic factors explaining anthocyanin pigmentation differences. BMC Plant Biol. 2024;24:627. https://doi.org/10.1186/s12870-024-05316-w.

35. Pucker B, Walker-Hale N, Dzurlic J, Yim WC, Cushman JC, Crum A, et al. Multiple mechanisms explain loss of anthocyanins from betalain-pigmented Caryophyllales, including repeated wholesale loss of a key anthocyanidin synthesis enzyme. New Phytol. 2024;241:471–89. https://doi.org/10.1111/nph.19341.

36. Zufall RA, Rausher MD. Genetic changes associated with floral adaptation restrict future evolutionary potential. Nature. 2004;428:847–50. https://doi.org/10.1038/nature02489.

37. Del Valle JC, Alcalde-Eon C, Escribano-Bailón MT, Buide ML, Whittall JB, Narbona E. Stability of petal color polymorphism: the significance of anthocyanin accumulation in photosynthetic tissues. BMC Plant Biol. 2019;19:496. https://doi.org/10.1186/s12870-019-2082-6.

38. Thomas MM, Rudall PJ, Ellis AG, Savolainen V, Glover BJ. Development of a complex floral trait: The pollinator-attracting petal spots of the beetle daisy, Gorteria diffusa (Asteraceae). Am J Bot. 2009;96:2184–96. https://doi.org/10.3732/ajb.0900079.

39. Fairnie ALM, Yeo MTS, Gatti S, Chan E, Travaglia V, Walker JF, et al. Eco-Evo-Devo of petal pigmentation patterning. Essays Biochem. 2022;66:753–68. https://doi.org/10.1042/EBC20220051.

40. Martins TR, Berg JJ, Blinka S, Rausher MD, Baum DA. Precise spatio-temporal regulation of the anthocyanin biosynthetic pathway leads to petal spot formation in Clarkia gracilis (Onagraceae). New Phytol. 2013;197:958–69. https://doi.org/10.1111/nph.12062.

41. Wang Z, Li X, Chen M, Yang L, Zhang Y. Molecular and Metabolic Insights into Anthocyanin Biosynthesis for Spot Formation on Lilium leichtlinii var. maximowiczii Flower Petals. Int J Mol Sci. 2023;24:1844. https://doi.org/10.3390/ijms24031844.

42. Ding B, Patterson EL, Holalu SV, Li J, Johnson GA, Stanley LE, et al. Two MYB Proteins in a Self-Organizing Activator-Inhibitor System Produce Spotted Pigmentation Patterns. Curr Biol CB. 2020;30:802-814.e8. https://doi.org/10.1016/j.cub.2019.12.067.

43. Zhang Y, Cheng Y, Ya H, Xu S, Han J. Transcriptome sequencing of purple petal spot region in tree peony reveals differentially expressed anthocyanin structural genes. Front Plant Sci. 2015;6.

44. Yuan Y-W, Rebocho AB, Sagawa JM, Stanley LE, Bradshaw HD. Competition between anthocyanin and flavonol biosynthesis produces spatial pattern variation of floral pigments between Mimulus species. Proc Natl Acad Sci. 2016;113:2448–53. https://doi.org/10.1073/pnas.1515294113.

45. Lin R-C, Rausher MD. R2R3-MYB genes control petal pigmentation patterning in Clarkia gracilis ssp. sonomensis (Onagraceae). New Phytol. 2021;229:1147–62. https://doi.org/10.1111/nph.16908.

46. Sasaki K, Takahashi T. A flavonoid from Brassica rapa flower as the UV-absorbing nectar guide. Phytochemistry. 2002;61:339–43. https://doi.org/10.1016/s0031-9422(02)00237-6.

47. Medel R, Botto-Mahan C, Kalin-Arroyo M. Pollinator-Mediated Selection on the Nectar Guide Phenotype in the Andean Monkey Flower, Mimulus luteus. Ecology. 2003;84:1721–32.

48. Lunau K, Fieselmann G, Heuschen B, van de Loo A. Visual targeting of components of floral colour patterns in flower-naïve bumblebees (Bombus terrestris; Apidae). Naturwissenschaften. 2006;93:325–8. https://doi.org/10.1007/s00114-006-0105-2.

49. Davies KM, Albert NW, Schwinn KE, Davies KM, Albert NW, Schwinn KE. From landing lights to mimicry: the molecular regulation of flower colouration and mechanisms for pigmentation patterning. Funct Plant Biol. 2012;39:619–38. https://doi.org/10.1071/FP12195.

50. Rudall PJ. Weird and wonderful foxgloves. 2016. https://www.kew.org/read-and-watch/weird-and-wonderful-foxgloves. Accessed 2 May 2024.

51. Keeble F, Pellew MC, Jones WN. The Inheritance Op Peloria and Flowercolour in Foxgloves (Digitalis purpurea). New Phytol. 1910;9:68–77. https://doi.org/10.1111/j.1469-8137.1910.tb05554.x.

52. Bradley D, Carpenter R, Copsey L, Vincent C, Rothstein S, Coen E. Control of inflorescence architecture in Antirrhinum. Nature. 1996;379:791–7. https://doi.org/10.1038/379791a0.

53. Shannon S, Meeks-Wagner DR. A Mutation in the Arabidopsis TFL1 Gene Affects Inflorescence Meristem Development. Plant Cell. 1991;3:877–92. https://doi.org/10.2307/3869152.

54. Bradley D, Ratcliffe O, Vincent C, Carpenter R, Coen E. Inflorescence Commitment and Architecture in Arabidopsis. Science. 1997;275:80–3. https://doi.org/10.1126/science.275.5296.80.

55. Alvarez J, Guli CL, Yu X-H, Smyth DR. terminal flower: a gene affecting inflorescence development in Arabidopsis thaliana. Plant J. 1992;2:103–16. https://doi.org/10.1111/j.1365-313X.1992.00103.x.

56. Pnueli L, Gutfinger T, Hareven D, Ben-Naim O, Ron N, Adir N, et al. Tomato SP-Interacting Proteins Define a Conserved Signaling System That Regulates Shoot Architecture and Flowering. Plant Cell. 2001;13:2687–702. https://doi.org/10.1105/tpc.010293.

57. Siadjeu C, Pucker B, Viehöver P, Albach DC, Weisshaar B. High Contiguity de novo Genome Sequence Assembly of Trifoliate Yam (Dioscorea dumetorum) Using Long Read Sequencing. Genes. 2020;11:274. https://doi.org/10.3390/genes11030274.

58. Wolff K, Friedhoff R, Schwarzer F, Pucker B. Data literacy in genome research. J Integr Bioinforma. 2023;20. https://doi.org/10.1515/jib-2023-0033.

59. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100. https://doi.org/10.1093/bioinformatics/bty191.

60. GrandOmics. NextDenovo. 2023.

61. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol. 2020;38:1044–53. https://doi.org/10.1038/s41587-020-0503-6.

62. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–6. https://doi.org/10.1038/s41587-019-0072-8.

63. GrandOmics. NextPolish. 2023.

64. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9. https://doi.org/10.1093/bioinformatics/btp352.

65. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021;38:4647–54. https://doi.org/10.1093/molbev/msab199.

66. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 2019;47:D807–11. https://doi.org/10.1093/nar/gky1053.

67. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34 suppl_2:W435–9. https://doi.org/10.1093/nar/gkl200.

68. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinforma Oxf Engl. 2011;27:757–63. https://doi.org/10.1093/bioinformatics/btr010.

69. Li H. lh3/yak. 2020.

70. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21:245. https://doi.org/10.1186/s13059-020-02134-9.

71. NCBI. sra-tools. 2020.

72. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21. https://doi.org/10.1093/bioinformatics/bts635.

73. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinforma. 2015;51:11.14.1-11.14.19. https://doi.org/10.1002/0471250953.bi1114s51.

74. Haak M, Vinke S, Keller W, Droste J, Rückert C, Kalinowski J, et al. High Quality de Novo Transcriptome Assembly of Croton tiglium. Front Mol Biosci. 2018;5. https://doi.org/10.3389/fmolb.2018.00062.

75. Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. 2016;44:e89. https://doi.org/10.1093/nar/gkw092.

76. Keilwagen J, Hartung F, Grau J. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. Methods Mol Biol Clifton NJ. 2019;1962:161–77. https://doi.org/10.1007/978-1-4939-9173-0_9.

77. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2. https://doi.org/10.1093/bioinformatics/btv351.

78. Pucker B, Holtgräwe D, Weisshaar B. Consideration of non-canonical splice sites improves gene prediction on the Arabidopsis thaliana Niederzenz-1 genome sequence. BMC Res Notes. 2017;10:667. https://doi.org/10.1186/s13104-017-2985-y.

79. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

80. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.

81. Pucker B, Iorizzo M. Apiaceae FNS I originated from F3H through tandem gene duplication. PLOS ONE. 2023;18:e0280155. https://doi.org/10.1371/journal.pone.0280155.

82. Rempel A, Choudhary N, Pucker B. KIPEs3: Automatic annotation of biosynthesis pathways. PLOS ONE. 2023;18:e0294342. https://doi.org/10.1371/journal.pone.0294342.

83. Pucker B. Automatic identification and annotation of MYB gene family members in plants. BMC Genomics. 2022;23:220. https://doi.org/10.1186/s12864-022-08452-5.

84. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30:772–80. https://doi.org/10.1093/molbev/mst010.

85. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE. 2010;5:e9490. https://doi.org/10.1371/journal.pone.0009490.

86. Thoben C, Pucker B. Automatic annotation of the bHLH gene family in plants. BMC Genomics. 2023;24:780. https://doi.org/10.1186/s12864-023-09877-2.

87. Edgar RC. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nat Commun. 2022;13:6968. https://doi.org/10.1038/s41467-022-34630-w.

88. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci. 2020;117:9451–7. https://doi.org/10.1073/pnas.1921046117.

89. Smit A, Hubley R, Green P. RepeatMasker. 2015.

90. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20:275. https://doi.org/10.1186/s13059-019-1905-y.

91. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 2000;16:276–7. https://doi.org/10.1016/S0168-9525(00)02024-2.

92. Eddy SR. Accelerated Profile HMM Searches. PLOS Comput Biol. 2011;7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

93. Paysan-Lafosse T, Andreeva A, Blum M, Chuguransky SR, Grego T, Pinto BL, et al. The Pfam protein families database: embracing AI/ML. Nucleic Acids Res. 2025;53:D523–34. https://doi.org/10.1093/nar/gkae997.

94. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35:W265–8. https://doi.org/10.1093/nar/gkm286.

95. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033.

96. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16:111–20. https://doi.org/10.1007/BF01731581.

97. Mokhtar MM, Alsamman AM, El Allali A. PlantLTRdb: An interactive database for 195 plant species LTR-retrotransposons. Front Plant Sci. 2023;14. https://doi.org/10.3389/fpls.2023.1134627.

98. González LG, Deyholos MK. Identification, characterization and distribution of transposable elements in the flax (Linum usitatissimum L.) genome. BMC Genomics. 2012;13:644. https://doi.org/10.1186/1471-2164-13-644.

99. Koch MA, 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. 2000;17:1483–98. https://doi.org/10.1093/oxfordjournals.molbev.a026248.

100. Gaut BS, Morton BR, McCaig BC, Clegg MT. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc Natl Acad Sci. 1996;93:10274–9. https://doi.org/10.1073/pnas.93.19.10274.

101. Ohshima K. RNA-Mediated Gene Duplication and Retroposons: Retrogenes, LINEs, SINEs, and Sequence Specificity. Int J Evol Biol. 2013;2013:424726. https://doi.org/10.1155/2013/424726.

102. Tang H, Krishnakumar V, Zeng X, Xu Z, Taranto A, Lomas JS, et al. JCVI: A versatile toolkit for comparative genomics analysis. iMeta. 2024;3:e211. https://doi.org/10.1002/imt2.211.

103. Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, et al. ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. Biochem Biophys Res Commun. 2012;419:779–81. https://doi.org/10.1016/j.bbrc.2012.02.101.

104. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.

105. Horz J. jahorz/Digitalis_1. 2026. https://doi.org/10.5281/zenodo.18637460.

106. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.

107. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab008. https://doi.org/10.1093/gigascience/giab008.

108. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7. https://doi.org/10.1038/nbt.3519.

109. Robinson JT, Thorvaldsdottir H, Turner D, Mesirov JP. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). Bioinformatics. 2023;39:btac830. https://doi.org/10.1093/bioinformatics/btac830.

110. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21:487–93. https://doi.org/10.1101/gr.113985.110.

111. Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B. An Arabidopsis thaliana T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. Plant Mol Biol. 2003;53:247–59. https://doi.org/10.1023/B:PLAN.0000009297.37235.4a.

112. Pucker B, Friedhoff R, Horz J, Wolff K. Genome sequence and annotation of Digitalis purpurea. 2023. https://doi.org/10.24355/dbbs.084-202311130950-0.

113. Horz J, Wolff K, Friedhoff R, Pucker B. Digitalis purpurea genome sequence and annotation V2. 2025. https://doi.org/10.60507/FK2/4KUKXI.

114. Tiley GP, Barker MS, Burleigh JG. Assessing the Performance of Ks Plots for Detecting Ancient Whole Genome Duplications. Genome Biol Evol. 2018;10:2882–98. https://doi.org/10.1093/gbe/evy200.

115. Richards MP. A brief review of the archaeological evidence for Palaeolithic and Neolithic subsistence. Eur J Clin Nutr. 2002;56:1270–8. https://doi.org/10.1038/sj.ejcn.1601646.

116. Forkmann G, Heller W, Grisebach H. Anthocyanin Biosynthesis in Flowers of Matthiola incana Flavanone 3-and Flavonoid 3′-Hydroxylases. Z Für Naturforschung C. 1980;35:691–5. https://doi.org/10.1515/znc-1980-9-1004.

117. Ferrer J-L, Jez JM, Bowman ME, Dixon RA, Noel JP. Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis. Nat Struct Biol. 1999;6:775–84. https://doi.org/10.1038/11553.

118. Jez JM, Bowman ME, Dixon RA, Noel JP. Structure and mechanism of the evolutionarily unique plant enzyme chalcone isomerase. Nat Struct Biol. 2000;7:786–91. https://doi.org/10.1038/79025.

119. Brugliera F, Barri-Rewell G, Holton TA, Mason JG. Isolation and characterization of a flavonoid 3′-hydroxylase cDNA clone corresponding to the Ht1 locus of Petunia hybrida. Plant J. 2002;19:441–51. https://doi.org/10.1046/j.1365-313X.1999.00539.x.

120. Saito K, Yonekura-Sakakibara K, Nakabayashi R, Higashi Y, Yamazaki M, Tohge T, et al. The flavonoid biosynthetic pathway in Arabidopsis: Structural and genetic diversity. Plant Physiol Biochem. 2013;72:21–34. https://doi.org/10.1016/j.plaphy.2013.02.001.

121. Abrahams S, Lee E, Walker AR, Tanner GJ, Larkin PJ, Ashton AR. The Arabidopsis TDS4 gene encodes leucoanthocyanidin dioxygenase (LDOX) and is essential for proanthocyanidin synthesis and vacuole development. Plant J. 2003;35:624–36. https://doi.org/10.1046/j.1365-313X.2003.01834.x.

122. Kim S, Binzel ML, Yoo KS, Park S, Pike LM. Pink (P), a new locus responsible for a pink trait in onions (Allium cepa) resulting from natural mutations of anthocyanidin synthase. Mol Genet Genomics. 2004;272:18–27. https://doi.org/10.1007/s00438-004-1041-5.

123. Jiang T, Zhang M, Wen C, Xie X, Tian W, Wen S, et al. Integrated metabolomic and transcriptomic analysis of the anthocyanin regulatory networks in Salvia miltiorrhiza Bge. flowers. BMC Plant Biol. 2020;20:349. https://doi.org/10.1186/s12870-020-02553-7.

124. Owen CR, Bradshaw HD. Induced mutations affecting pollinator choice in Mimulus lewisii (Phrymaceae). Arthropod-Plant Interact. 2011;5:235–44. https://doi.org/10.1007/s11829-011-9133-8.

125. Abid MA, Wei Y, Meng Z, Wang Y, Ye Y, Wang Y, et al. Increasing floral visitation and hybrid seed production mediated by beauty mark in Gossypium hirsutum. Plant Biotechnol J. 2022;20:1274–84. https://doi.org/10.1111/pbi.13805.

126. Richter R, Dietz A, Foster J, Spaethe J, Stöckl A. Flower patterns improve foraging efficiency in bumblebees by guiding approach flight and landing. Funct Ecol. 2023;37:763–77. https://doi.org/10.1111/1365-2435.14262.

127. Johnson SD, Midgley JJ. Fly pollination of Gorteria diffusa (Asteraceae), and a possible mimetic function for dark spots on the capitulum. Am J Bot. 1997;84:429–36. https://doi.org/10.2307/2446018.

128. de Jager ML, Willis-Jones E, Critchley S, Glover BJ. The impact of floral spot and ring markings on pollinator foraging dynamics. Evol Ecol. 2017;31:193–204. https://doi.org/10.1007/s10682-016-9852-5.