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

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

1 Plant Biotechnology and Bioinformatics, Institute of Plant Biology & BRICS, TU Braunschweig, Mendelssohnstrasse 4, 38106 Braunschweig, Germany

* corresponding author: b.pucker@tu-braunschweig.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 and the producer of the biomedical compound digoxin. Here, we present a long read sequencing-based genome sequence of a red 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 red 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. In addition, the anthocyanin biosynthesis activator *MYB5* shows a 18 bp deletion in white flowering plants that results in the loss of 6 amino acids in the protein. Furthermore, we found a large insertion in the *DpTFL1/CEN* gene to be 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 (Kreis, 2017). 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 (Bucca, 2018; Carroll *et al.*, 2023). The plant contains a variety of cardiac glycosides, with the most important being digoxin (Bucca, 2018). These compounds work by increasing the force and efficiency of heart contractions, which are important to treat heart failure and other cardiac conditions. *D. purpurea* and the close relative *D. lanata* are cultivated for the extraction of cardenolides (Kreis, 2017). 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. Many experiments explored the potential of tissue cultures to propagate plants and to improve the yield of medicinal compound production (Kreis, 2017). A cytochrome P450 CYP87A4 is crucial for the biosynthesis of digoxin (Carroll *et al.*, 2023). The identification of all genes required for a biosynthesis process is the first step towards production in a heterologous system.

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 purple or pink 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 (Winkel-Shirley, 2001; Pucker & Selmar, 2022). 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 (Winkel-Shirley, 2001; Pucker *et al.*, 2020) (**Fig.1c**). A recent study revealed anthocyanin-related glutathione S-transferase (arGST) as an additional enzyme in the anthocyanin biosynthesis (Eichenberger *et al.*, 2023). 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 (Goodman *et al.*, 2004; Zhao & Dixon, 2009; Pucker & Selmar, 2022). A previously

postulated ‘ligandin’, that might protect anthocyanins during the transport through the cytoplasm (Mueller *et al.*, 2000), was recently characterized as an enzyme in the cyanidin biosynthesis, which makes an involvement in the anthocyanin transport less likely (Eichenberger *et al.*, 2023). Besides a direct transport through the cytoplasm, there are several studies that report an involvement of vesicles in the anthocyanin transport in different plant species (Nozue & Yasuda, 1985; Grotewold *et al.*, 1998; Zhang *et al.*, 2006; Poustka *et al.*, 2007; Pucker & Selmar, 2022).

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 (Luo *et al.*, 2016). Subtle differences in substrate specificity of the competing enzymes e.g. flavonol synthase (FLS) and DFR can determine the pattern of synthesized flavonoids (Johnson *et al.*, 2001; Choudhary & Pucker, 2024). 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 (Choudhary & Pucker, 2024).

Activity of different branches of the flavonoid biosynthesis is mostly controlled at the transcriptional level through regulation of gene expression by transcription factors (Weisshaar & Jenkins, 1998; Li, 2014; Xu *et al.*, 2015; Lloyd *et al.*, 2017). 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 (Ramsay & Glover, 2005; Gonzalez *et al.*, 2008; Lloyd *et al.*, 2017). 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 (Feller *et al.*, 2011; Albert *et al.*, 2021). The WD40 component is often TTG1, but LWD1 and LWD2 can also participate in complex formation (Airoldi *et al.*, 2019). An additional WRKY component (TTG2) has been suggested to specifically increase the expression of genes associated with the anthocyanin transport into the vacuole (Gonzalez *et al.*, 2016; Lloyd *et al.*, 2017). 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* (Borevitz *et al.*, 2000; Tohge *et al.*, 2005; Gonzalez *et al.*, 2008). MYB5 was identified as an anthocyanin regulator in *Fragaria* (Jiang *et al.*, 2023), *Vitis* (Cavallini *et al.*, 2014), *Mimulus* (Zheng *et al.*, 2021), and *Nelumbo* (Sun *et al.*, 2016). It was shown that MYB5 forms a MBW complex with LWR1/LWD1-like instead of TTG1 to activate the anthocyanin biosynthesis (Jiang *et al.*, 2023). A study in *Vaccinium* suggests an involvement of the proanthocyanidin regulator MYB123 in the anthocyanin biosynthesis (Karppinen *et al.*, 2021). While most anthocyanin regulating MYBs are activators, the incorporation of subgroup 4 MYBs into the MBW complex can result in an inhibitory function (Albert *et al.*, 2014).

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 (Quattrocchio *et al.*, 1999; Streisfeld *et al.*, 2013; Wheeler *et al.*, 2022; Marin-Recinos & Pucker,

2024). 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 (Marin-Recinos & Pucker, 2024). 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 (Pucker *et al.*, 2024). Changes in flower color often 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 (Zufall & Rausher, 2004). 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 (Del Valle *et al.*, 2019).

The flower pigmentation of *D. purpurea* is not uniform, but characterized by intensely pigmented dark spots on the bottom petals (**Fig. 1c**). Petal spots have been reported in other plant species including *Gorteria diffusa* (Thomas *et al.*, 2009), *Kohleria warszewiczii* (Fairnie *et al.*, 2022), *Clarkia gracilis* (Martins *et al.*, 2013), *Lilium leichtlinii* (Wang *et al.*, 2023), and *Mimulus guttatus* (Ding *et al.*, 2020). Pigmentation spots on petals can be caused by the accumulation of high levels of anthocyanins (cyanidin derivatives) (Martins *et al.*, 2013; Zhang *et al.*, 2015). These spots can be surrounded by a white region characterized by high levels of flavonols, the products of a pathway competing with anthocyanin biosynthesis (Yuan *et al.*, 2016). 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) (Ding *et al.*, 2020). 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 (Ding *et al.*, 2020). RTO proteins appear to travel through the endosomal system to other cells, where they act as repressors (Ding *et al.*, 2020). Another study in *C. gracilis* reported an anthocyanin activating MYB (CgsMYB12) as the causal gene for the formation of a single large spot (Lin & Rausher, 2021) 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 (Sasaki & Takahashi, 2002; Medel *et al.*, 2003; Lunau *et al.*, 2006; Davies *et al.*, 2012; Fairnie *et al.*, 2022).

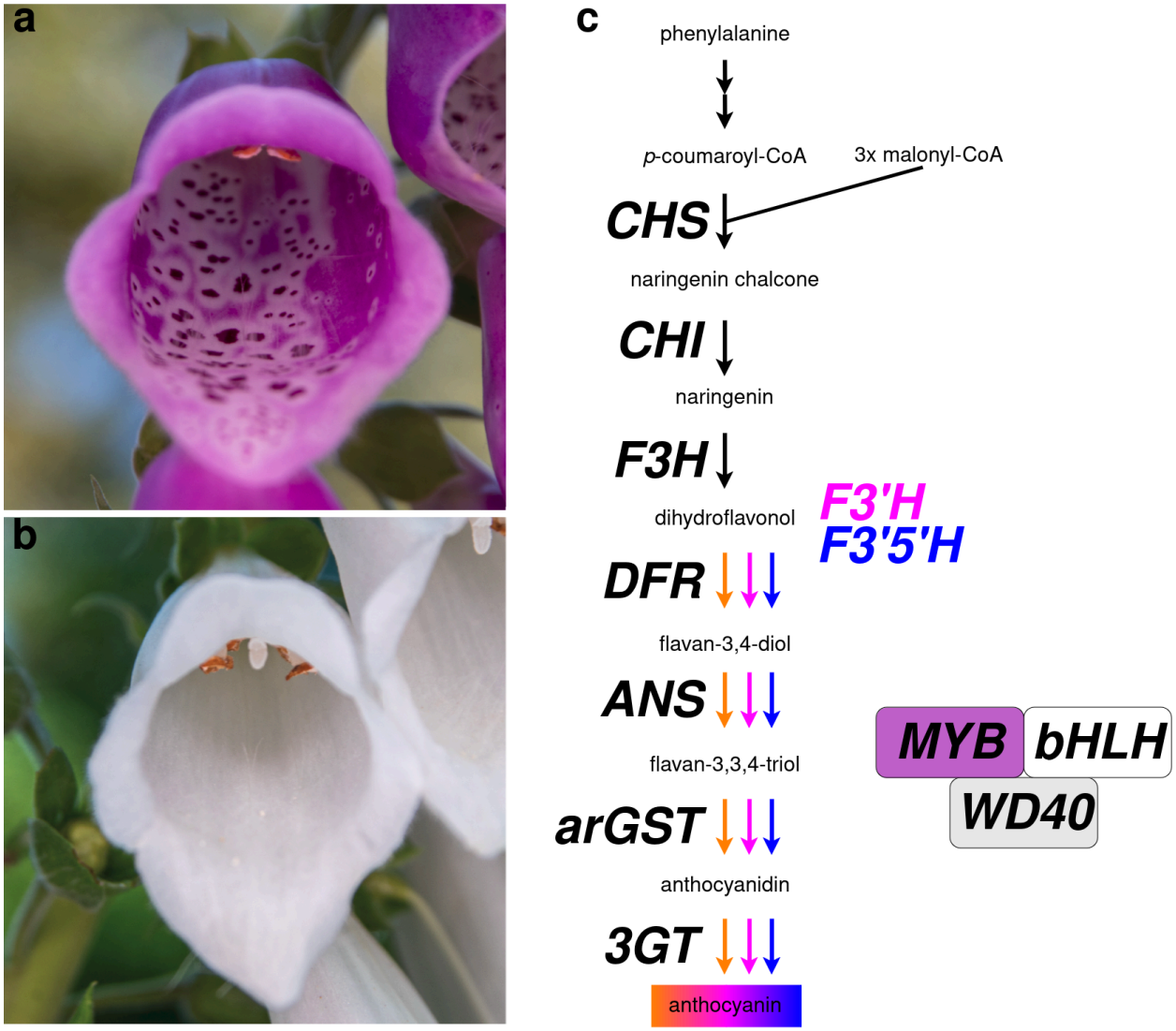


**Fig.1**: Pictures showing red (a) and white (b) flowers of *Digitalis purpurea* (foxglove). The anthocyanin biosynthesis is displayed as the central branch of the flavonoid biosynthesis (c). Abbreviations: CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-hydroxylase; DFR, dihydroflavonol 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 (Rudall, 2016). 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

(Keeble *et al.*, 1910). The characterization of the genetics behind the development of a terminal flower in *D. purpurea* was not investigated yet. The occurrence of terminal flowers in plants with otherwise indeterminate inflorescence architecture has been described in different species like *Antirrhinum* (Bradley *et al.*, 1996) and Arabidopsis (Shannon & Meeks-Wagner, 1991). In *Antirrhinum*, a loss of function mutation in the *CEN* (*CENTRORADIALIS*) gene leads to the development of a terminal flower (Bradley *et al.*, 1996). The functional CEN gene is expressed in the subapical meristem and disrupts the initiation of flower development (Bradley *et al.*, 1996). Therefore, CEN is a flowering repressor. The Arabidopsis ortholog TERMINAL FLOWERING 1 (TFL1) plays a similar role (Bradley *et al.*, 1997). 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 (Shannon & Meeks-Wagner, 1991; Alvarez *et al.*, 1992). 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 (Pnueli *et al.*, 2001). They have been shown to interact directly with other proteins so they might be part of a more elaborate signaling system (Pnueli *et al.*, 2001).

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

Here, we reported 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 red flowering individuals revealed molecular differences in the flavonoid biosynthesis genes that could explain the pigmentation loss in the white flowering plants. The molecular basis of large terminal flowers 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 red flowering plant in a garden close to these coordinates 51.92351 N, 8.84441 E. 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 (Siadjeu *et al.*, 2020). 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 (Wolff *et al.*, 2023) 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 (Li, 2018) 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 red flowering plant. NextDenovo v2.5.0 (GrandOmics, 2023a), Shasta 0.10.0 (Shafin *et al.*, 2020), and Flye 2.9.1-b1780 (Kolmogorov *et al.*, 2019) 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 (GrandOmics, 2023b) based on read mappings that were generated with minimap2 (Li, 2018) and samtools v1.15.1 (Li *et al.*, 2009).

## Structural annotation

Paired-end RNA-seq reads (SRR128113,SRR128114,SRR128115,ERR2040595,Additional file 1) were retrieved from the Sequence Read Archive via fastq-dump (NCBI, 2020) and aligned to the assembled *D. purpurea* genome sequence with STAR v2.5.1b (Dobin *et al.*, 2013; Dobin & Gingeras, 2015) in 2-pass-mode using a minimal similarity of 95% and a minimal alignment length of 90% as previously described (Haak *et al.*, 2018) to provide hints for the gene prediction process. Gene models were predicted with BRAKER v3.0.2 (Gabriel *et al.*, 2023) based on the *Arabidopsis thaliana* species parameters and GeneMark v4 (Brůna *et al.*, 2020). BUSCO v3.0.2 (Simão *et al.*, 2015; Manni *et al.*, 2021) was used to assess the completeness of the genome sequence and the completeness of the gene prediction with BRAKER. The genomic BUSCO assessment command contained the arguments '--long --limit 10 -sp arabidopsis', $10^{-3}$ as e-value cutoff for BLAST, and eudicotyledon_odb10 (embryophyta) as the reference. AUGUSTUS v3.2.2 (Stanke *et al.*, 2006; Keller *et al.*, 2011) was used for the gene prediction. HMMER v3.1b2 (Finn *et al.*, 2011; Mistry *et al.*, 2013; Eddy, 2023) was run to confirm orthology of candidates. The entire process was repeated after a species specific training for optimization of the gene prediction. When running BUSCO in proteome mode, the same configuration was used except for the gene prediction arguments. All predicted protein encoding genes were checked for expression evidence to differentiate between active genes and potentially inactive pseudogenes. A cutoff of 1 TPM in any of the analyzed RNA-seq datasets (Additional file 1) was defined to distinguish active genes from potentially inactive pseudogenes.

## 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 (Pucker *et al.*, 2017; Haak *et al.*, 2018) 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+ (Altschul *et al.*, 1990, 1997) 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 (Pucker & Iorizzo, 2023). Enzyme and transporter encoding genes of the flavonoid biosynthesis were identified using KIPEs v3.2.2 with the flavonoid biosynthesis data set v3 (Pucker *et al.*, 2020; Rempel *et al.*, 2023) with additional parameters listed in Additional file 2. Flavonoid biosynthesis controlling MYB transcription factors were annotated with the MYB_annotator v0.231(Pucker, 2022). The following parameters were used for the MYB annotation: initial candidate identification via BLASTp v2.13.0+ (Altschul *et al.*, 1990, 1997), alignment construction with MAFFT v7.475 (Katoh & Standley, 2013), phylogenetic tree construction with FastTree v2.1.10 (Price *et al.*, 2010), 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 v1.03 (Thoben & Pucker, 2023). The following parameters were used for the bHLH annotation: identification of initial candidates via BLASTp v2.13.0+ (Altschul *et al.*, 1990, 1997), alignment with muscle v5.1 (Edgar, 2022), phylogenetic tree construction with FastTree v2 (Price *et al.*, 2010), 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).

## Analysis of gene expression

Fastq-dump (NCBI, 2020) 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 (Bray *et al.*, 2016) 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 (Pucker & Iorizzo, 2023; Choudhary & Pucker, 2024).

## Identification of sequence variants associated with anthocyanin loss

Reads of two red flowering plants and two white flowering plants (Additional file 5) were aligned against the reference genome sequence of a red flowering plant with minimap v2.24-r1122 (Li, 2018). The resulting BAM files were indexed with samtools v1.13 (Li *et al.*, 2009). Candidate genes associated with the flavonoid biosynthesis were manually inspected for systematic sequence differences via Integrative Genomics Viewer (IGV) v2.15.4 (Robinson *et al.*, 2023).

## Genotyping of a *Digitalis purpurea* population

A population of 89 plants was subjected to genotyping via PCR for the discovered mutations responsible for 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 beds located at 52.2804716735375 N and 10.548701669439103 E.

DNA for subsequent genotyping PCR was extracted based on the GABI-Kat CTAB DNA preparation protocol (Rosso *et al.*, 2003). This protocol, designed for *A. thaliana,* was adapted for the use on *D. purpurea*: Only one young leaf was harvested from each plant. After extraction the DNA samples were stored at -20 °C.

PCR was conducted with oligonucleotides flanking the *ANS* and the *TFL1*/*CEN* loci, respectively (Additional file 10, Additional file 11). 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 red colored petioles (Additional file 12). These phenotypes were later validated by adding information about the flower color.

# Results

## Genome sequence assembly and structural annotation

A red flowering plant (DR1) was selected for sequencing and construction of a reference genome sequence (Pucker *et al.*, 2023). The rationale behind this selection was the assumption that a red flowering plant should have all genes required for the biosynthesis of anthocyanins which are most likely responsible for the red flower pigmentation. For comparison, another red 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.

**Table 1**: Comparison of assembler performance for the analysis of *Digitalis purpurea* long read sequencing data of the red 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 initially revealed 329,325 protein encoding genes (DR1_v1a1; (Pucker *et al.*, 2023). A completeness check revealed 93% complete BUSCO genes in this annotation. The presence of 44% duplicated BUSCO genes suggested that a large fraction of the assembly represents separate haplotypes. This was supported by a coverage analysis that found an average coverage of 20x for large parts of the assembly, while small parts showed a coverage of around 40x (Additional file 6). The analysis of transcript abundances based on RNA-seq data revealed 56,065 active genes (TPM > 1 in at least one sample), while 273,260 predicted genes showed close to zero evidence for expression in any of the studied samples (Additional file 7).

## 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.3**; Additional file 8). Additionally, the transcription factors controlling the anthocyanin biosynthesis genes in *D. purpurea* were investigated. MYB genes placed in the *MYB5* lineage (g25977, g268730, g7955,g166017), in the *MYB75*/*PAP1* lineage (g147607), and in the *MYB123/TT2* lineage (g26269) were identified. Candidates for the bHLH partner in the anthocyanin biosynthesis-regulating MBW complex (*bHLH42/TT8*) were identified in *D. purpurea*: g55107.t1 and g55108.t1. Additionally, orthologs of the WD40 genes *TTG1* (g109077) and *LWD1* (g183911,g251558), and for the bHLH MYC4 (g166063, g268685) were identified in a phylogenetic analysis. In summary, all genes expected for a complete anthocyanin biosynthesis pathway were successfully identified in the reference genome sequence DR1_v1.

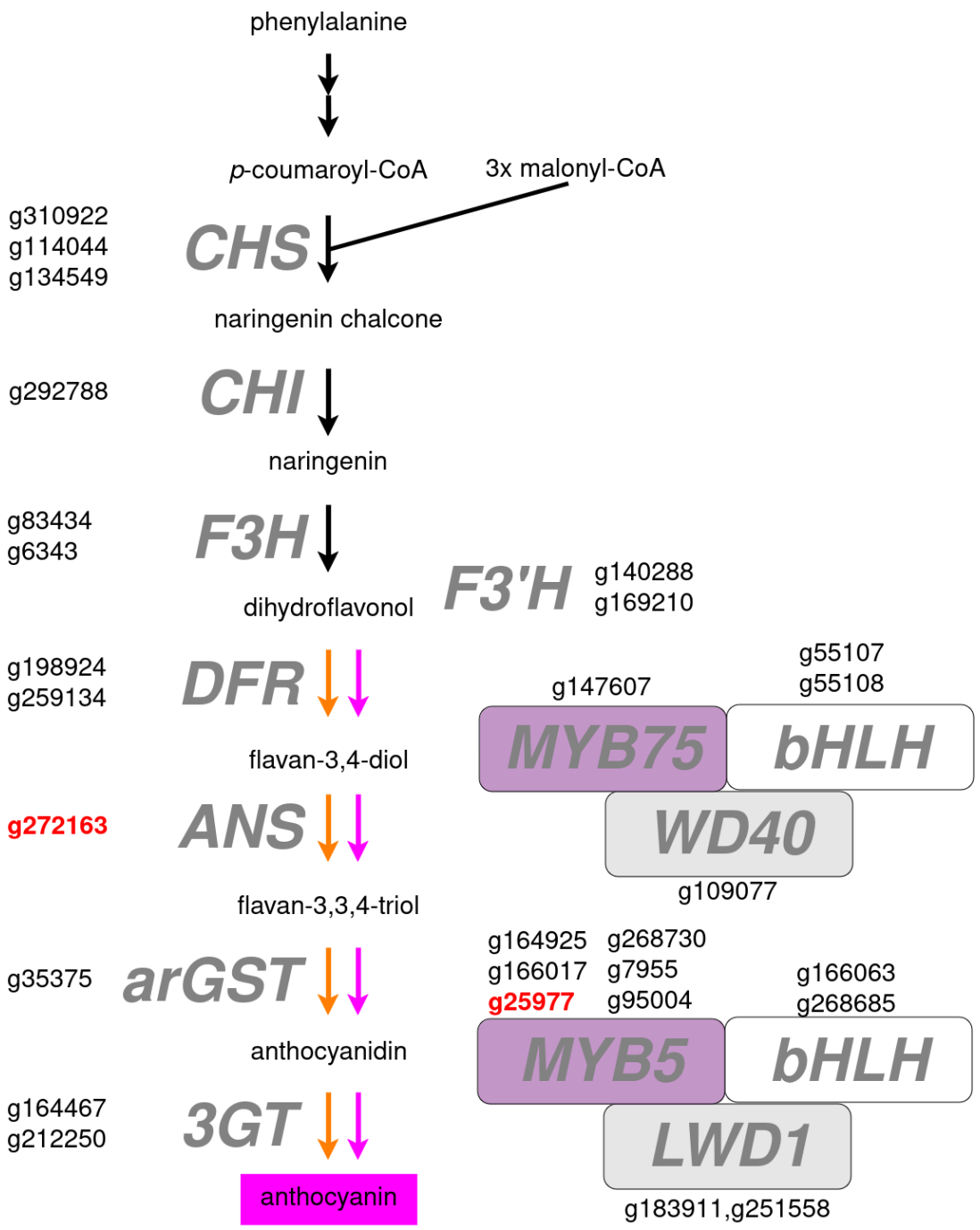


**Fig.3**: *Digitalis purpurea* candidate genes associated with the anthocyanin biosynthesis including the transcriptional regulators of the MBW complex. Yellow and magenta arrows indicate that the biosynthesis of pelargonidin and cyanidin derivatives would be possible, while the lack of a flavonoid 3',5'-dihydroxylase (*F3'5'H*) does not allow the synthesis of delphinidin derivatives. *CHS*, naringenin-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; MYB, Myeloblastosis; bHLH, basic helix-loop-helix; TTG1, TRANSPARENT TESTA GLABRA 1; LWD1, LIGHT-REGULATED WD1.

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). 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-pigment plant organs.

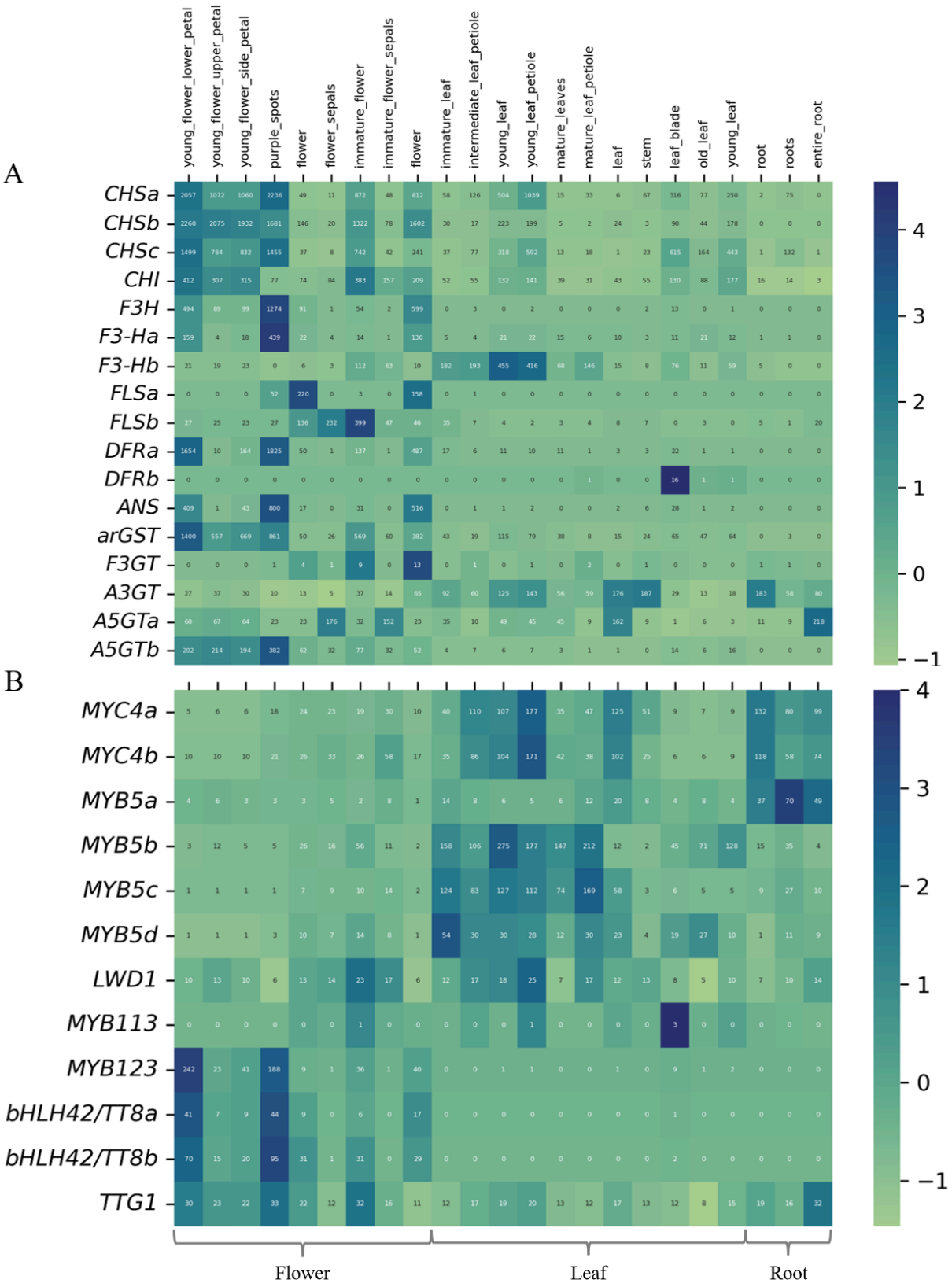
A
young_flower_lower_petal
young_flower_upper_petal
young_flower_side_petal
purple_spots
flower
flower_sepals
immature_flower
immature_flower_sepals
flower
immature_leaf
intermediate_leaf_petiole
young_leaf
young_leaf_petiole
mature_leaves
mature_leaf_petiole
leaf
stem
leaf_blade
old_leaf
young_leaf
root
roots
entire_root
CHSa
CHSb
CHSc
CHI
F3H
F3-Ha
F3-Hb
FLSa
FLSb
DFRa
DFRb
ANS
arGST
F3GT
A3GT
A5GTa
A5GTb
B
MYC4a
MYC4b
MYB5a
MYB5b
MYB5c
MYB5d
LWD1
MYB113
MYB123
bHLH42/TT8a
bHLH42/TT8b
TTG1
Flower
Leaf
Root
4
3
2
1
0
−1

**Fig.4**: Expression plot showing the activity of anthocyanin biosynthesis associated genes across a range of different plant parts. (A) Expression rates of genes encoding pathway enzymes. (B) Expression rates of possible anthocyanin related transcription factors. 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; *FLS*, flavonol synthase, *DFR*, dihydroflavonol 4-reductase; *ANS*, anthocyanidin synthase; *arGST*, anthocyanin-related glutathione S-transferase; *F3GT*, UDP-glucose flavonoid 3-O-glucosyltransferase; *A3GT*, UDP-glucose anthocyanidin 3-O-glucosyltransferase; *A5GT*, UDP-glucose anthocyanidin 5-O-glucosyltransferase; *MYC*, bHLH-type MYC; MYB, Myeloblastosis; bHLH, basic helix-loop-helix; *TTG1*, *TRANSPARENT TESTA GLABRA 1*; *LWD1*, *LIGHT-REGULATED WD1*.

## Sequence variants underlying pigmentation differences

All identified candidate genes associated with the anthocyanin biosynthesis were manually inspected in a read mapping comparing red flowering and white flowering plants. The most striking difference is an insertion of approximately 12 kb large into the *ANS* gene. This insertion is homozygous in the two white flowering plants, while the red flowering plants harbor an *ANS* allele without this insertion. Since the reference genome sequence DR1_v1 contains the 12 kb sequence, the automatic prediction of the *ANS* gene model is inaccurate. The alignment of RNA-seq reads indicates that *ANS* is not located at the end of this insertion and downstream of it as suggested by the gene model, but instead spans the entire insertion with one exon being located upstream (**Fig.5**).

Genotyping of 89 plants for the *ANS* mutation marker identified 14 plants with the homozygous wildtype allele, 31 plants with a heterozygous genotype, and 44 plants with the homozygous mutant allele. Unfortunately, only a part of the population has flowered so far. For all flowering plants the presence or absence of anthocyanins in the plant has been accurately derived from the leaf stem by visual inspection as confirmed by the flower colors. For 14 plants, flower color phenotypes were identified. These phenotypes correlate perfectly with the plants' genotypes thus supporting *ans* as the causal locus in white flowering *D. purpurea* plants.

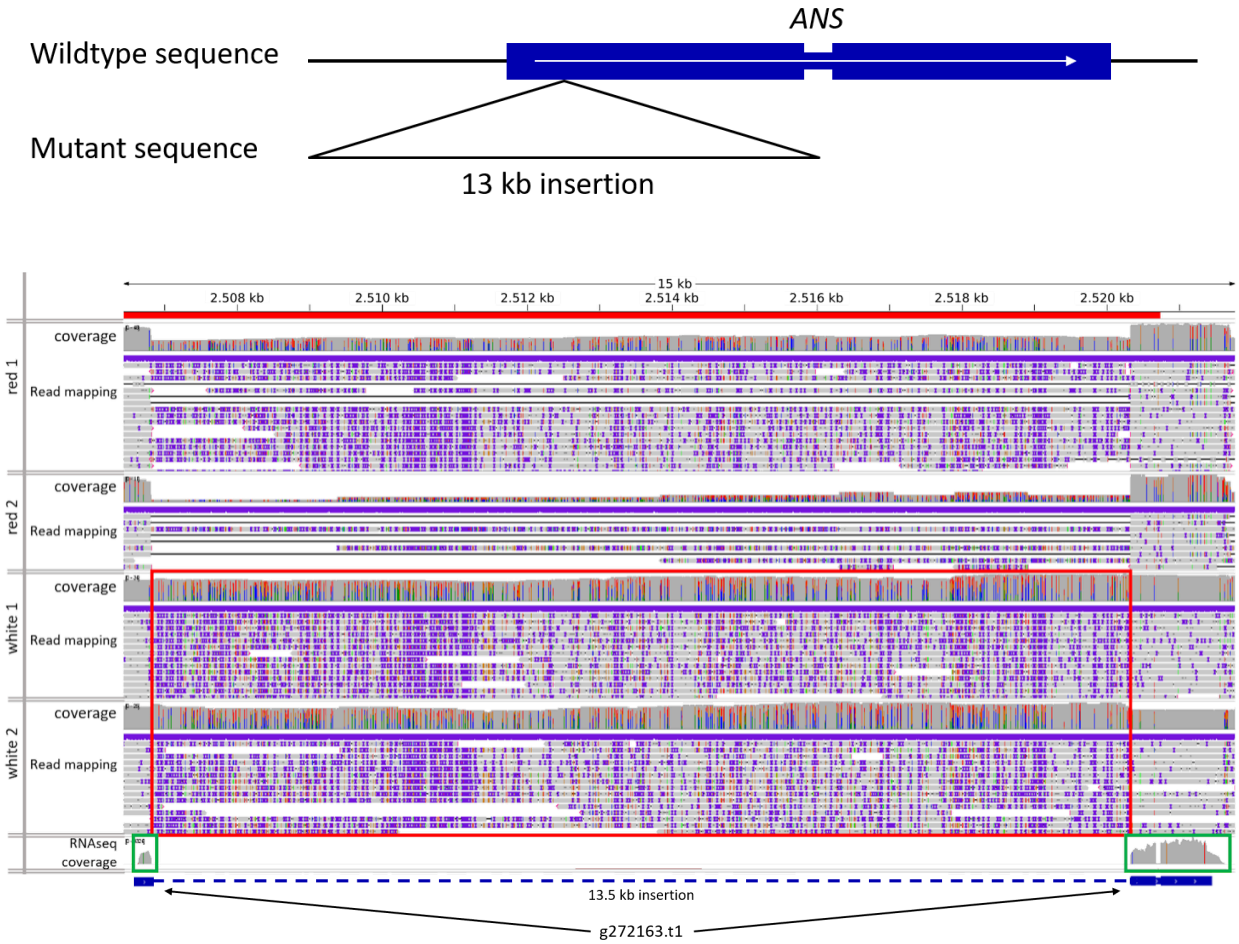


**Fig.5**: Comparison of the *ANS* locus in red and white flowering *D. purpurea* plants revealed a large insertion that appears homozygous in the white flowering plants. The insertion is expected to be a loss-of-function mutation that results in the loss of ANS enzymatic activity and thus a block in the anthocyanin biosynthesis ultimately resulting in the white flower phenotype.

Another systematic difference between the red group and the white group is located inside the MYB5 candidate g25977 (*DpMYB5*). This gene showed a homozygous 18 bp deletion in the sequencing reads of the white flowering plants (**Fig. 6**). At least some reads of the red flowering plants contain this 18 bp deletion thus suggesting that the mutation cannot be homozygous in the red flowering plants. The 18 bp deletion leads to the lack of the amino acids “KKVKET” at position 152 to 157 in the DpMYB5 protein. Structural analysis with alphafold2 revealed the location where the difference between the mutant and the wildtype protein is located.This loss of six amino acids does not affect the well conserved MYB domain, but might influence the transcriptional activation ability of this protein.

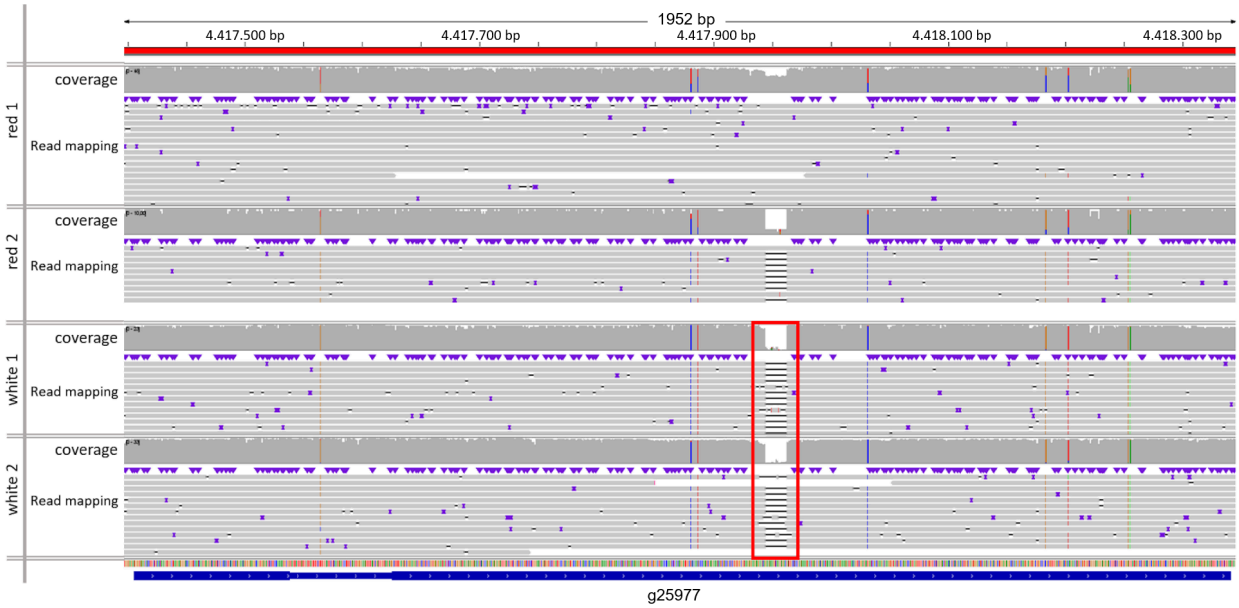


**Fig.6**: Genomic region showing the deletion of 18 bp in the coding sequence of *DpMYB5* (g25977) in long reads belonging to white flowering plants, while reads to red flowering plants show that this region is present.

## 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 phenotypical plant were compared to three wildtype plants. Upon investigation of the *Digitalis* orthologs of the *Arabidopsis TFL1* a 14337 bp insertion was found starting only 4 base pairs after the start codon of the gene and leading to a premature stop codon after 4 codons (**Fig. 7**). The insertion has a high similarity to the insertion in the *ans* gene described above.

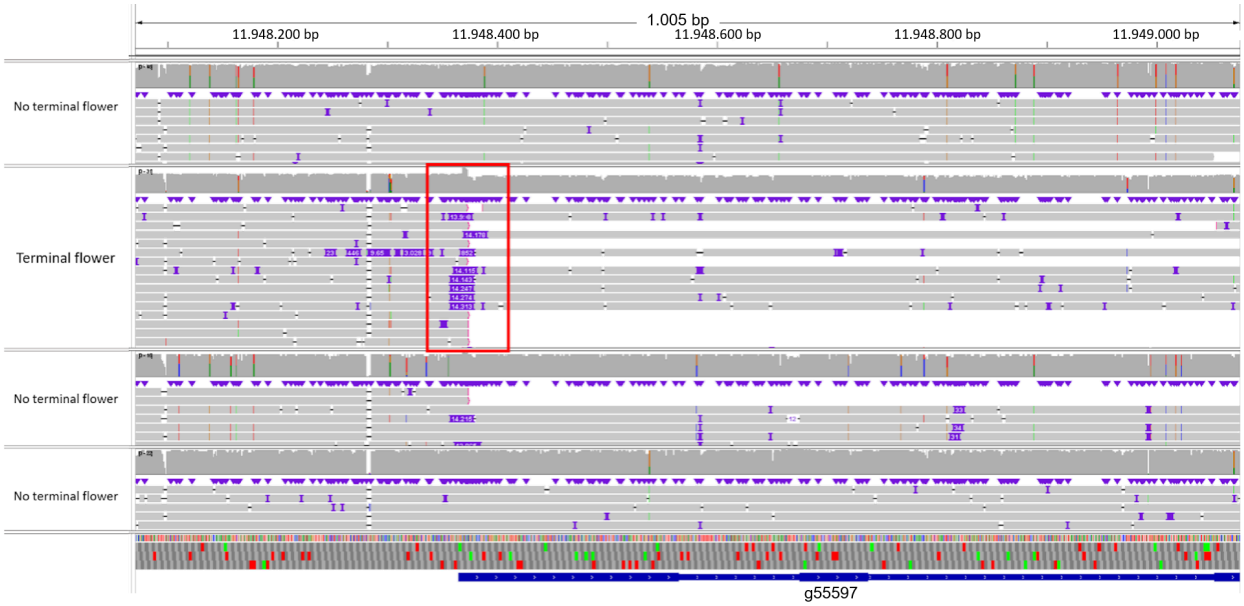


**Fig. 7**. Genomic region showing the insertion of 14 kb in the *DpFL1/CEN* (g55597) in the reads of the plant with a terminal flower (red box). The plants with indeterminate flowering architecture show at least some reads without the insert.

# Discussion

The highly continuous genome sequence of a red flowering *D. purpurea* and the corresponding annotation of protein encoding genes enabled the identification of genes associated with the anthocyanin biosynthesis. The original number of 273,260 predicted protein encoding genes was reduced to the more realistic number of 56,065 genes by considering only genes that show solid transcript evidence in any of the available RNA-seq datasets. Similar approaches have been discussed and applied before to reduce large numbers of predicted gene models based on transcription evidence to the most important ones (Liang *et al.*, 2009; Dohm *et al.*, 2014; McGrath *et al.*, 2022). Given that structural annotation workflows do not provide perfectly accurate results, evaluation and filtering is recommended (Vuruputoor *et al.*, 2023). Although RNA-seq samples from a range of different plant organs and treatments were included in this filtering based on transcript evidence, it is possible that transcripts of *bona fide* genes were not detected in the RNA-seq data leading to a mis-classification of the corresponding genes as non-expressed pseudogene. According to previous reports, only 60-70% of all encoded genes are typically expressed in a given sample (Salojärvi *et al.*, 2017). Given that plant genomes harbor large numbers of genes that are activated in response to stress conditions, it seems obvious that many genes are only active under specific conditions.

Long read sequencing of a white flowering individual 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 the two white flowering plants, while the two red 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*.
We noticed that flavonols are produced in the white flowering plants (Additional file 9). 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 (Forkmann *et al.*, 1980; Ferrer *et al.*, 1999; Jez *et al.*, 2000; Brugliera *et al.*, 2002; Saito *et al.*, 2013). 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 identified 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 (Abrahams *et al.*, 2003; Kim *et al.*, 2004; Jiang *et al.*, 2020). While the above described *ans* mutation could be a universal explanation for the lack of pigments in white *D. purpurea* flowers. Theoretically, it remains possible that different mechanisms are responsible for the lack of pigmentation in different white flowering *D. purpurea* individuals.

As transcription factors are often implicated in the loss of anthocyanin pigmentation (Marin-Recinos & Pucker, 2024), the identified candidate genes were checked for sequence variants associated with the anthocyanin loss. An ortholog of the previously described anthocyanin biosynthesis regulator MYB5 (Jiang *et al.*, 2023) was discovered to show a 18 bp deletion in *D. purpurea*. While this sequence variant could disrupt the activator domain of this transcription factor, the low expression of this gene in flower samples suggest that *DpMYB5* might only control the anthocyanin formation plant parts except for the flower. However, an analysis with genetic markers suggest that *DpMYB5* is at least genetically linked to the locus responsible for the flower pigmentation. Expression of *DpMYB5* in the leaves suggests that this MYB might trigger the anthocyanin biosynthesis in response to light stress.

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 (Owen & Bradshaw, 2011; Abid *et al.*, 2022; Richter *et al.*, 2023). 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 (Johnson & Midgley, 1997; de Jager *et al.*, 2017). This suggests the assumption that the pigmentation patterns in *D. purpurea* flowers 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 9**)**.

The present study focuses on explaining the completely white flowering phenotype (**Fig.1**), but there are also cases of white flowers with red spots in *D. purpurea*. Similar patterns have been described before in other plant species (Ding *et al.*, 2020). 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 (Bradley *et al.*, 1996, 1997). The system responsible for inflorescence architecture seems to be highly conserved among plants (Pnueli *et al.*, 2001), suggesting a similar system in *D. purpurea* as well. Genotyping of a *D. purpurea* population for the *DpTFL1/CEN* locus revealed that only the homozygous mutant genotype leads to the development of a determinate inflorescence architecture, corroborating the findings of Keeble *et al*. (Keeble *et al.*, 1910).

# 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 LeoPARD: https://doi.org/10.24355/dbbs.084-202311130950-0 (Pucker *et al.*, 2023).

## Competing interests

The authors declare that they have no competing interests.

## Funding

We acknowledge support by the Open Access Publication Funds of Technische Universität Braunschweig.

## Authors' contributions

BP planned the study. KW, RF, and JH 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 BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). 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.

## Additional files

**Additional file 1**: *Digitalis purpurea* RNA-seq data sets that delivered hints for the gene prediction process.
**Additional file 2**: *Digitalis purpurea* long read sequencing runs with corresponding accessions assigned by the European Nucleotide Archive (ENA).
**Additional file 3**: Documentation of all parameters and input files used for the annotation of the flavonoid biosynthesis genes with KIPEs3.
**Additional file 4**: Documentation of all parameters and input files used for the annotation of the MYB transcription factor gene family with the MYB_annotator.
**Additional file 5**: Documentation of all parameters and input files used for the annotation of the bHLH transcription factor gene family with the bHLH_annotator.
**Additional file 6**: *Digitalis purpurea* candidate genes associated with the anthocyanin biosynthesis.
**Additional file 7**: Prediction of pseudogenes based on the gene expression (TPM<1) across all available samples.
**Additional file 8**: 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 9**: Purple 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.
**Additional file 10**: Design of the genotyping PCRs for the *ANS* and *TFL1/CEN* locus.
**Additional file 11**: PCR cycle program for the *DpCEN* and *DpANS* genotyping.
**Additional file 12**: *D. purpurea* plants showing red colored petioles.

## References


**Abid MA, Wei Y, Meng Z, Wang Y, Ye Y, Wang Y, He H, Zhou Q, Li Y, Wang P, *et al.*2022**. Increasing floral visitation and hybrid seed production mediated by beauty mark in Gossypium hirsutum. *Plant Biotechnology Journal* **20**: 1274–1284.
**Abrahams S, Lee E, Walker AR, Tanner GJ, Larkin PJ, Ashton AR**. **2003**. The Arabidopsis TDS4 gene encodes leucoanthocyanidin dioxygenase (LDOX) and is essential for proanthocyanidin synthesis and vacuole development. *The Plant Journal* **35**: 624–636.
**Airoldi CA, Hearn TJ, Brockington SF, Webb AAR, Glover BJ**. **2019**. TTG1 proteins regulate circadian activity as well as epidermal cell fate and pigmentation. *Nature Plants* **5**: 1145–1153.
**Albert NW, Butelli E, Moss SMA, Piazza P, Waite CN, Schwinn KE, Davies KM, Martin C**. **2021**. Discrete bHLH transcription factors play functionally overlapping roles in pigmentation patterning in flowers of Antirrhinum majus. *New Phytologist* **231**: 849–863.
**Albert NW, Davies KM, Lewis DH, Zhang H, Montefiori M, Brendolise C, Boase MR, Ngo H, Jameson PE, Schwinn KE**. **2014**. A conserved network of transcriptional activators and repressors regulates anthocyanin pigmentation in eudicots. *The Plant Cell* **26**: 962–980.
**Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ**. **1990**. Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.
**Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ**. **1997**. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.
**Alvarez J, Guli CL, Yu X-H, Smyth DR**. **1992**. terminal flower: a gene affecting inflorescence development in Arabidopsis thaliana. *The Plant Journal* **2**: 103–116.
**Borevitz JO, Xia Y, Blount J, Dixon RA, Lamb C**. **2000**. Activation Tagging Identifies a Conserved MYB Regulator of Phenylpropanoid Biosynthesis. *The Plant Cell* **12**: 2383–2393.
**Bradley D, Carpenter R, Copsey L, Vincent C, Rothstein S, Coen E**. **1996**. Control of inflorescence architecture in Antirrhinum. *Nature* **379**: 791–797.
**Bradley D, Ratcliffe O, Vincent C, Carpenter R, Coen E**. **1997**. Inflorescence Commitment and Architecture in Arabidopsis. *Science* **275**: 80–83.
**Bray NL, Pimentel H, Melsted P, Pachter L**. **2016**. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**: 525–527.
**Brugliera F, Barri-Rewell G, Holton TA, Mason JG**. **2002**. Isolation and characterization of a flavonoid 3′-hydroxylase cDNA clone corresponding to the Ht1 locus of Petunia hybrida. *The Plant Journal* **19**: 441–451.
**Brůna T, Lomsadze A, Borodovsky M**. **2020**. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics* **2**.
**Bucca A**. **2018**. Role of Digoxin in Heart Failure. In: Vasan RS, Sawyer DB, eds. Encyclopedia of Cardiovascular Research and Medicine. Oxford: Elsevier, 323–326.
**Carroll E, Ravi Gopal B, Raghavan I, Mukherjee M, Wang ZQ**. **2023**. A cytochrome P450 CYP87A4 imparts sterol side-chain cleavage in digoxin biosynthesis. *Nature Communications* **14**: 4042.
**Cavallini E, Zenoni S, Finezzo L, Guzzo F, Zamboni A, Avesani L, Tornielli GB**. **2014**. Functional Diversification of Grapevine MYB5a and MYB5b in the Control of Flavonoid

Biosynthesis in a Petunia Anthocyanin Regulatory Mutant. *Plant and Cell Physiology* **55**: 517–534.
**Choudhary N, Pucker B**. **2024**. Conserved amino acid residues and gene expression patterns associated with the substrate preferences of the competing enzymes FLS and DFR. : 2023.11.05.565693.
**Davies KM, Albert NW, Schwinn KE, Davies KM, Albert NW, Schwinn KE**. **2012**. From landing lights to mimicry: the molecular regulation of flower colouration and mechanisms for pigmentation patterning. *Functional Plant Biology* **39**: 619–638.
**Del Valle JC, Alcalde-Eon C, Escribano-Bailón MT, Buide ML, Whittall JB, Narbona E**. **2019**. Stability of petal color polymorphism: the significance of anthocyanin accumulation in photosynthetic tissues. *BMC Plant Biology* **19**: 496.
**Ding B, Patterson EL, Holalu SV, Li J, Johnson GA, Stanley LE, Greenlee AB, Peng F, Bradshaw HD, Blinov ML, *et al.*2020**. Two MYB Proteins in a Self-Organizing Activator-Inhibitor System Produce Spotted Pigmentation Patterns. *Current biology: CB* **30**: 802-814.e8.
**Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR**. **2013**. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.
**Dobin A, Gingeras TR**. **2015**. Mapping RNA-seq Reads with STAR. *Current Protocols in Bioinformatics* **51**: 11.14.1-11.14.19.
**Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, Rupp O, Sörensen TR, Stracke R, Reinhardt R, *et al.*2014**. The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). *Nature* **505**: 546–549.
**Eddy SR**. **2023**. HMMER User's Guide.
**Edgar RC**. **2022**. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications* **13**: 6968.
**Eichenberger M, Schwander T, Hüppi S, Kreuzer J, Mittl PRE, Peccati F, Jiménez-Osés G, Naesby M, Buller RM**. **2023**. The catalytic role of glutathione transferases in heterologous anthocyanin biosynthesis. *Nature Catalysis* **6**: 927–938.
**Fairnie ALM, Yeo MTS, Gatti S, Chan E, Travaglia V, Walker JF, Moyroud E**. **2022**. Eco-Evo-Devo of petal pigmentation patterning. *Essays in Biochemistry* **66**: 753–768.
**Feller A, Machemer K, Braun EL, Grotewold E**. **2011**. Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. *The Plant Journal* **66**: 94–116.
**Ferrer J-L, Jez JM, Bowman ME, Dixon RA, Noel JP**. **1999**. Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis. *Nature Structural Biology* **6**: 775–784.
**Finn RD, Clements J, Eddy SR**. **2011**. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**: W29–W37.
**Forkmann G, Heller W, Grisebach H**. **1980**. Anthocyanin Biosynthesis in Flowers of Matthiola incana Flavanone 3-and Flavonoid 3′-Hydroxylases. *Zeitschrift für Naturforschung C* **35**: 691–695.
**Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, Stanke M**. **2023**. BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. : 2023.06.10.544449.
**Gonzalez A, Brown M, Hatlestad G, Akhavan N, Smith T, Hembd A, Moore J, Montes D, Mosley T, Resendez J, *et al.*2016**. TTG2 controls the developmental regulation of seed coat tannins in Arabidopsis by regulating vacuolar transport steps in the proanthocyanidin pathway. *Developmental Biology* **419**: 54–63.
**Gonzalez A, Zhao M, Leavitt JM, Lloyd AM**. **2008**. Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in Arabidopsis seedlings. *The Plant Journal* **53**: 814–827.
**Goodman CD, Casati P, Walbot V**. **2004**. A Multidrug Resistance–Associated Protein Involved in Anthocyanin Transport in Zea mays. *The Plant Cell* **16**: 1812–1826.

**GrandOmics**. **2023a**. NextDenovo.
**GrandOmics**. **2023b**. NextPolish.
**Grotewold E, Chamberlin M, Snook M, Siame B, Butler L, Swenson J, Maddock S, St. Clair G, Bowen B**. **1998**. Engineering Secondary Metabolism in Maize Cells by Ectopic Expression of Transcription Factors. *The Plant Cell* **10**: 721–740.
**Haak M, Vinke S, Keller W, Droste J, Rückert C, Kalinowski J, Pucker B**. **2018**. High Quality de Novo Transcriptome Assembly of Croton tiglium. *Frontiers in Molecular Biosciences* **5**.
**de Jager ML, Willis-Jones E, Critchley S, Glover BJ**. **2017**. The impact of floral spot and ring markings on pollinator foraging dynamics. *Evolutionary Ecology* **31**: 193–204.
**Jez JM, Bowman ME, Dixon RA, Noel JP**. **2000**. Structure and mechanism of the evolutionarily unique plant enzyme chalcone isomerase. *Nature Structural Biology* **7**: 786–791.
**Jiang L, Yue M, Liu Y, Zhang N, Lin Y, Zhang Y, Wang Y, Li M, Luo Y, Zhang Y, *et al.*2023**. A novel R2R3-MYB transcription factor FaMYB5 positively regulates anthocyanin and proanthocyanidin biosynthesis in cultivated strawberries (Fragaria × ananassa). *Plant Biotechnology Journal* **21**: 1140–1158.
**Jiang T, Zhang M, Wen C, Xie X, Tian W, Wen S, Lu R, Liu L**. **2020**. Integrated metabolomic and transcriptomic analysis of the anthocyanin regulatory networks in Salvia miltiorrhiza Bge. flowers. *BMC Plant Biology* **20**: 349.
**Johnson SD, Midgley JJ**. **1997**. Fly pollination of Gorteria diffusa (Asteraceae), and a possible mimetic function for dark spots on the capitulum. *American Journal of Botany* **84**: 429–436.
**Johnson ET, Ryu S, Yi H, Shin B, Cheong H, Choi G**. **2001**. Alteration of a single amino acid changes the substrate specificity of dihydroflavonol 4-reductase. *The Plant Journal: For Cell and Molecular Biology* **25**: 325–333.
**Karppinen K, Lafferty DJ, Albert NW, Mikkola N, McGhie T, Allan AC, Afzal BM, Häggman H, Espley RV, Jaakola L**. **2021**. MYBA and MYBPA transcription factors co-regulate anthocyanin biosynthesis in blue-coloured berries. *New Phytologist* **232**: 1350–1367.
**Katoh K, Standley DM**. **2013**. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**: 772–780.
**Keeble F, Pellew MC, Jones WN**. **1910**. The Inheritance Op Peloria and Flowercolour in Foxgloves (digitalis Purpurea). *New Phytologist* **9**: 68–77.
**Keller O, Kollmar M, Stanke M, Waack S**. **2011**. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics (Oxford, England)* **27**: 757–763.
**Kim S, Binzel ML, Yoo KS, Park S, Pike LM**. **2004**. Pink (P), a new locus responsible for a pink trait in onions (Allium cepa) resulting from natural mutations of anthocyanidin synthase. *Molecular Genetics and Genomics* **272**: 18–27.
**Kolmogorov M, Yuan J, Lin Y, Pevzner PA**. **2019**. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37**: 540–546.
**Kreis W**. **2017**. The Foxgloves (Digitalis) Revisited. *Planta Medica* **83**: 962–976.
**Li S**. **2014**. Transcriptional control of flavonoid biosynthesis. *Plant Signaling & Behavior* **9**: e27522.
**Li H**. **2018**. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.
**Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup**. **2009**. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
**Liang C, Mao L, Ware D, Stein L**. **2009**. Evidence-based gene predictions in plant genomes. *Genome Research* **19**: 1912–1923.
**Lin R-C, Rausher MD**. **2021**. R2R3-MYB genes control petal pigmentation patterning in Clarkia gracilis ssp. sonomensis (Onagraceae). *New Phytologist* **229**: 1147–1162.
**Lloyd A, Brockman A, Aguirre L, Campbell A, Bean A, Cantero A, Gonzalez A**. **2017**. 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 and Cell Physiology* **58**: 1431–1441.
**Lunau K, Fieselmann G, Heuschen B, van de Loo A**. **2006**. Visual targeting of components of floral colour patterns in flower-naïve bumblebees (Bombus terrestris; Apidae). *Naturwissenschaften* **93**: 325–328.
**Luo P, Ning G, Wang Z, Shen Y, Jin H, Li P, Huang S, Zhao J, Bao M**. **2016**. Disequilibrium of Flavonol Synthase and Dihydroflavonol-4-Reductase Expression Associated Tightly to White vs. Red Color Flower Formation in Plants. *Frontiers in Plant Science* **6**.
**Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM**. **2021**. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **38**: 4647–4654.
**Marin-Recinos MF, Pucker B**. **2024**. Genetic factors explaining anthocyanin pigmentation differences. *BMC Plant Biology* **24**: 627.
**Martins TR, Berg JJ, Blinka S, Rausher MD, Baum DA**. **2013**. Precise spatio-temporal regulation of the anthocyanin biosynthetic pathway leads to petal spot formation in Clarkia gracilis (Onagraceae). *New Phytologist* **197**: 958–969.
**McGrath JM, Funk A, Galewski P, Ou S, Townsend B, Davenport K, Daligault H, Johnson S, Lee J, Hastie A, *et al.*2022**. A contiguous de novo genome assembly of sugar beet EL10 (Beta vulgaris L.). *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* **30**: dsac033.
**Medel R, Botto-Mahan C, Kalin-Arroyo M**. **2003**. Pollinator-Mediated Selection on the Nectar Guide Phenotype in the Andean Monkey Flower, Mimulus luteus. *Ecology* **84**: 1721–1732.
**Mistry J, Finn RD, Eddy SR, Bateman A, Punta M**. **2013**. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research* **41**: e121.
**Mueller LA, Goodman CD, Silady RA, Walbot V**. **2000**. AN9, a Petunia Glutathione S-Transferase Required for Anthocyanin Sequestration, Is a Flavonoid-Binding Protein. *Plant Physiology* **123**: 1561–1570.
**NCBI**. **2020**. sra-tools.
**Nozue M, Yasuda H**. **1985**. Occurrence of anthocyanoplasts in cell suspension cultures of sweet potato. *Plant Cell Reports* **4**: 252–255.
**Owen CR, Bradshaw HD**. **2011**. Induced mutations affecting pollinator choice in Mimulus lewisii (Phrymaceae). *Arthropod-Plant Interactions* **5**: 235–244.
**Pnueli L, Gutfinger T, Hareven D, Ben-Naim O, Ron N, Adir N, Lifschitz E**. **2001**. Tomato SP-Interacting Proteins Define a Conserved Signaling System That Regulates Shoot Architecture and Flowering. *The Plant Cell* **13**: 2687–2702.
**Poustka F, Irani NG, Feller A, Lu Y, Pourcel L, Frame K, Grotewold E**. **2007**. 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 Physiology* **145**: 1323–1335.
**Price MN, Dehal PS, Arkin AP**. **2010**. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**: e9490.
**Pucker B**. **2022**. Automatic identification and annotation of MYB gene family members in plants. *BMC Genomics* **23**: 220.
**Pucker B, Friedhoff R, Horz J, Wolff K**. **2023**. Genome sequence and annotation of Digitalis purpurea.
**Pucker B, Holtgräwe D, Weisshaar B**. **2017**. Consideration of non-canonical splice sites improves gene prediction on the Arabidopsis thaliana Niederzenz-1 genome sequence. *BMC Research Notes* **10**: 667.
**Pucker B, Iorizzo M**. **2023**. Apiaceae FNS I originated from F3H through tandem gene duplication. *PLOS ONE* **18**: e0280155.
**Pucker B, Reiher F, Schilbert HM**. **2020**. Automatic Identification of Players in the Flavonoid

Biosynthesis with Application on the Biomedicinal Plant Croton tiglium. *Plants* **9**: 1103.
**Pucker B, Selmar D**. **2022**. Biochemistry and Molecular Basis of Intracellular Flavonoid Transport in Plants. *Plants (Basel, Switzerland)* **11**: 963.
**Pucker B, Walker-Hale N, Dzurlic J, Yim WC, Cushman JC, Crum A, Yang Y, Brockington SF**. **2024**. Multiple mechanisms explain loss of anthocyanins from betalain-pigmented Caryophyllales, including repeated wholesale loss of a key anthocyanidin synthesis enzyme. *New Phytologist* **241**: 471–489.
**Quattrocchio F, Wing J, Woude K van der, Souer E, Vetten N de, Mol J, Koes R**. **1999**. Molecular Analysis of the anthocyanin2 Gene of Petunia and Its Role in the Evolution of Flower Color. *The Plant Cell* **11**: 1433–1444.
**Ramsay NA, Glover BJ**. **2005**. MYB-bHLH-WD40 protein complex and the evolution of cellular diversity. *Trends in Plant Science* **10**: 63–70.
**Rempel A, Choudhary N, Pucker B**. **2023**. KIPEs3: Automatic annotation of biosynthesis pathways. *PLOS ONE* **18**: e0294342.
**Richter R, Dietz A, Foster J, Spaethe J, Stöckl A**. **2023**. Flower patterns improve foraging efficiency in bumblebees by guiding approach flight and landing. *Functional Ecology* **37**: 763–777.
**Robinson JT, Thorvaldsdottir H, Turner D, Mesirov JP**. **2023**. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics* **39**: btac830.
**Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B**. **2003**. An Arabidopsis thaliana T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. *Plant Molecular Biology* **53**: 247–259.
**Rudall PJ**. **2016**. Weird and wonderful foxgloves.
**Saito K, Yonekura-Sakakibara K, Nakabayashi R, Higashi Y, Yamazaki M, Tohge T, Fernie AR**. **2013**. The flavonoid biosynthetic pathway in Arabidopsis: Structural and genetic diversity. *Plant Physiology and Biochemistry* **72**: 21–34.
**Salojärvi J, Smolander O-P, Nieminen K, Rajaraman S, Safronov O, Safdari P, Lamminmäki A, Immanen J, Lan T, Tanskanen J, *et al.*2017**. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature Genetics* **49**: 904–912.
**Sasaki K, Takahashi T**. **2002**. A flavonoid from Brassica rapa flower as the UV-absorbing nectar guide. *Phytochemistry* **61**: 339–343.
**Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, *et al.*2020**. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology* **38**: 1044–1053.
**Shannon S, Meeks-Wagner DR**. **1991**. A Mutation in the Arabidopsis TFL1 Gene Affects Inflorescence Meristem Development. *The Plant Cell* **3**: 877–892.
**Siadjeu C, Pucker B, Viehöver P, Albach DC, Weisshaar B**. **2020**. High Contiguity de novo Genome Sequence Assembly of Trifoliate Yam (Dioscorea dumetorum) Using Long Read Sequencing. *Genes* **11**: 274.
**Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM**. **2015**. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212.
**Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B**. **2006**. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**: W435–W439.
**Streisfeld MA, Young WN, Sobel JM**. **2013**. Divergent selection drives genetic differentiation in an R2R3-MYB transcription factor that contributes to incipient speciation in Mimulus aurantiacus. *PLoS genetics* **9**: e1003385.
**Sun S-S, Gugger PF, Wang Q-F, Chen J-M**. **2016**. Identification of a R2R3-MYB gene regulating anthocyanin biosynthesis and relationships between its variation and flower color

difference in lotus (Nelumbo Adans.). *PeerJ* **4**: e2369.
**Thoben C, Pucker B**. **2023**. Automatic annotation of the bHLH gene family in plants. *BMC Genomics* **24**: 780.
**Thomas MM, Rudall PJ, Ellis AG, Savolainen V, Glover BJ**. **2009**. Development of a complex floral trait: The pollinator-attracting petal spots of the beetle daisy, Gorteria diffusa (Asteraceae). *American Journal of Botany* **96**: 2184–2196.
**Tohge T, Nishiyama Y, Hirai MY, Yano M, Nakajima J, Awazuhara M, Inoue E, Takahashi H, Goodenowe DB, Kitayama M, *et al.*2005**. Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. *The Plant Journal: For Cell and Molecular Biology* **42**: 218–235.
**Vuruputoor VS, Monyak D, Fetter KC, Webster C, Bhattarai A, Shrestha B, Zaman S, Bennett J, McEvoy SL, Caballero M, *et al.*2023**. Welcome to the big leaves: Best practices for improving genome annotation in non-model plant genomes. *Applications in Plant Sciences* **11**: e11533.
**Wang Z, Li X, Chen M, Yang L, Zhang Y**. **2023**. Molecular and Metabolic Insights into Anthocyanin Biosynthesis for Spot Formation on Lilium leichtlinii var. maximowiczii Flower Petals. *International Journal of Molecular Sciences* **24**: 1844.
**Weisshaar B, Jenkins GI**. **1998**. Phenylpropanoid biosynthesis and its regulation. *Current Opinion in Plant Biology* **1**: 251–257.
**Wheeler LC, Walker JF, Ng J, Deanna R, Dunbar-Wallis A, Backes A, Pezzi PH, Palchetti MV, Robertson HM, Monaghan A, *et al.*2022**. Transcription Factors Evolve Faster Than Their Structural Gene Targets in the Flavonoid Pigment Pathway. *Molecular Biology and Evolution* **39**: msac044.
**Winkel-Shirley B**. **2001**. Flavonoid Biosynthesis. A Colorful Model for Genetics, Biochemistry, Cell Biology, and Biotechnology. *Plant Physiology* **126**: 485–493.
**Wolff K, Friedhoff R, Schwarzer F, Pucker B**. **2023**. Data literacy in genome research. *Journal of Integrative Bioinformatics* **20**.
**Xu W, Dubos C, Lepiniec L**. **2015**. Transcriptional control of flavonoid biosynthesis by MYB–bHLH–WDR complexes. *Trends in Plant Science* **20**: 176–185.
**Yuan Y-W, Rebocho AB, Sagawa JM, Stanley LE, Bradshaw HD**. **2016**. Competition between anthocyanin and flavonol biosynthesis produces spatial pattern variation of floral pigments between Mimulus species. *Proceedings of the National Academy of Sciences* **113**: 2448–2453.
**Zhang Y, Cheng Y, Ya H, Xu S, Han J**. **2015**. Transcriptome sequencing of purple petal spot region in tree peony reveals differentially expressed anthocyanin structural genes. *Frontiers in Plant Science* **6**.
**Zhang H, Wang L, Deroles S, Bennett R, Davies K**. **2006**. New insight into the structures and formation of anthocyanic vacuolar inclusions in flower petals. *BMC Plant Biology* **6**: 29.
**Zhao J, Dixon RA**. **2009**. MATE Transporters Facilitate Vacuolar Uptake of Epicatechin 3′-O-Glucoside for Proanthocyanidin Biosynthesis in Medicago truncatula and Arabidopsis. *The Plant Cell* **21**: 2323–2340.
**Zheng X, Om K, Stanton KA, Thomas D, Cheng PA, Eggert A, Simmons E, Yuan Y-W, Conradi Smith GD, Puzey JR, *et al.*2021**. The regulatory network for petal anthocyanin pigmentation is shaped by the MYB5a/NEGAN transcription factor in Mimulus. *Genetics* **217**.
**Zufall RA, Rausher MD**. **2004**. Genetic changes associated with floral adaptation restrict future evolutionary potential. *Nature* **428**: 847–850.