

Genome sequence of the medicinal plant *Tropaeolum majus* provides insights into flavonoid biosynthesis

Ronja Friedhoff¹, Ute Wittstock², Boas Pucker^{1*}

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

² Institute of Pharmaceutical Biology, Technische Universität Braunschweig, Mendelssohnstrasse 1, 38106 Braunschweig, Germany

* corresponding author: b.pucker@tu-braunschweig.de

RF: 0009-0002-0270-4661

UW: 0000-0002-0914-8453

BP: 0000-0002-3321-7471

Abstract

Tropaeolum majus is a very popular species around the world with an enormous number of commercially available varieties displaying various flower color patterns and growth characteristics. It is rich in phytochemicals such as glucotropaeolin, hydroxycinnamic acid derivatives, and flavonol glycosides. Here, we report a highly continuous genome sequence of *T. majus* and a comprehensive annotation of protein encoding genes suitable for comparative genomics. The potential for the exploration of individual gene functions in this valuable plant is demonstrated by an analysis of the flavonoid biosynthesis genes and their transcriptional regulators. Important players of the flavonol biosynthesis, including structural genes and a transcription factor, were identified that are required to produce precursors of phytomedically relevant flavonol glycosides. The genome sequence does not reveal an ortholog of the leucoanthocyanidin reductase encoding gene (*LAR*), which aligns with previous reports about the absence of this gene in many Brassicales species.

Introduction

The Tropaeolaceae family is native to South and Central America and represents one of the most ancient families of the Brassicales order. Its last common ancestor with the more recent core Brassicales has been dated to about 100 million years ago and its separation from Akaniaceae in the same clade to about 75 million years ago (Cardinal-McTeague *et al.*, 2016). According to molecular-phylogenetic analyses, the about 90 species of the Tropaeolaceae can all be assigned to one genus, *Tropaeolum* (nasturtium), with several sections (Andersson & Andersson, 2000). Plants of this family are fleshy vines characterized by their large, mostly monosymmetric flowers with a nectar spur and their twining petioles (Missouri Botanical Garden, 2024). Because of their eye-catching flowers, several species, including *Tropaeolum majus*, *T. peregrinum* and *T. speciosum*, are grown as ornamental plants in gardens and parks. *T. tuberosum* (mashua) is grown in the Andes for its tuberous roots which are consumed by humans as a rich source of starch, protein, and fibers (Acurio *et al.*, 2024).

T. majus is a very popular species around the world with an enormous number of commercially available varieties with various flower color patterns and growth characteristics. Its flowers are edible, have a spicy taste and are, for example, used to decorate salads. It is also used as a medicinal plant (Albrecht *et al.*, 2023). For example, a preparation composed of *T. majus* herbal powder and *Armoracia rusticana* root powder has been approved in Germany for treatment of acute inflammatory diseases such as respiratory tract and urinary infections. The entire plant is rich in phytochemicals such as glucotropaeolin, hydroxycinnamic acid derivatives, and flavonol glycosides (Česlová *et al.*, 2023). Apart from its horticultural and medicinal uses, *T. majus* has been subject of scientific research in various fields of plant biology. In the course of studies on cell wall biosynthesis, RNA-seq data have been generated from seeds and have been used to identify additional enzymes involved in xyloglucan biosynthesis (Jensen *et al.*, 2012). However, a genome sequence, which would enable in-depth studies of primary and specialized metabolite biosynthesis pathways in *T. majus*, their regulation and evolution and further breeding efforts of this and other important species of the genus, is still missing. Here, we report a highly continuous genome sequence of *T. majus* and an annotation of the protein encoding genes. This genomic data set enabled a systematic investigation of the flavonoid biosynthesis in this valuable plant.

Results & Discussion

Genome sequence and annotation of *T. majus*

A total of 97.7 Gbp of ONT sequencing data was produced. Multiple assembly attempts were conducted (**Table 1**). Finally, the NextDenovo assembly was selected as the final genome sequence due to higher contiguity and a higher completeness. No bacterial or fungal contamination contigs were detected in this assembly. This is also supported by the GC content of 35% that is close to the average observed for other plant genome sequences (Pucker & Brockington, 2018). With an N50 of 35 Mbp, the continuity of this assembly exceeds the values

reported for most previously released plant genome sequences (Marks *et al.*, 2021) and provides a solid basis for comparative genomics. Synteny analyses and the discovery of biosynthetic gene clusters are facilitated with the release of this genome sequence.

Table 1: Statistics of the *T. majus* genome sequence assemblies.

	NextDenovo	Shasta
Number of contigs	149	168
Total assembly size	872.9 Mbp	808.5 Mbp
Largest contig	69.5 Mbp	60.9 Mbp
N50	35 Mbp	26.2 Mbp
BUSCO v3	C:96.0%[S:92.0%,D:4.0%], F:0.6%,M:3.4%,n:2326	C:95.1%[S:91.1%,D:4.0%], F:0.7%,M:4.2%,n:2326

Multiple gene prediction approaches were conducted based on the assembled genome sequence (**Table 2**). The annotation produced by GeMoMa based on structures of protein coding genes in other plant species performed best based on BUSCO results and was selected as the final annotation. The number of 32,924 predicted protein encoding genes is close to the average observed in plants (Pucker & Brockington, 2018). A prediction of the gene functions was generated based on sequence similarity to *Arabidopsis thaliana* sequences and information available from TAIR10/Araport11 (Pucker *et al.*, 2024).

Table 2: Selected gene prediction results for the *T. majus* genome sequence. Different protein hints were used for: ¹ BRAKER (Viridiplantae) and ² GeMoMa (*Carica papaya*, *Theobroma cacao*, *Arabidopsis halleri*, *A. thaliana*, *Eutrema salsugineum*, and *Thlaspi arvense*).

	AUGUSTUS	BRAKER	GeMoMa
Hints	Ab initio	Protein hints ¹	Protein hints ²
Number of genes	107,093	185,031	32,924
Number of transcripts	107,093	193,038	52,836
BUSCO v5	C:83.0% [S:78.7%,D:4.3%], F:8.1%,M:8.9%, n:2326	C:77.3% [S:71.2%,D:6.1%], F:9.9%,M:12.8%, n:2326	C:96.8% [S:76.7%,D:20.1%], F:0.9%,M:2.3%, n:2326

Investigation of the flavonoid biosynthesis in *T. majus*

Genes for all steps of the core flavonoid biosynthesis were identified in the *T. majus* gene set (Table 3, Fig. 1). The presence of flavonoid biosynthesis genes in *T. majus* is expected given previous reports about the presence of flavonoids (Garzón *et al.*, 2015; Česlová *et al.*, 2023). Especially relevant for the biosynthesis of previously reported flavonol glycosides in *T. majus* (Garzón *et al.*, 2015; Česlová *et al.*, 2023) are the genes *CHS*, *CHI*, *F3H*, *F3'H*, *F3'5'H*, and *FLS* (Winkel-Shirley, 2001; Pucker *et al.*, 2020). Although not all land plants harbor *F3'5'H*, the detection in *T. majus* aligns with reports of myricetin derivatives (Garzón *et al.*, 2015) which requires hydroxylation reactions that can be catalyzed by *F3'5'H*. A gene encoding a member of the well known flavonol biosynthesis regulating R2R3-MYB subgroup 7 (Stracke *et al.*, 2007, 2010) was identified which could be a target for future attempts to engineer the flavonol biosynthesis in this medicinal plant.

No gene encoding leucoanthocyanidin reductase (LAR) was detected in this *T. majus* gene set. Given the phylogenetic position of *T. majus* within the Brassicales, this aligns with previous reports about the absence of LAR in the Brassicales (Nesi *et al.*, 2009; Pucker *et al.*, 2020; Lu, 2024). The functional and ecological consequences of the absence of LAR remain unknown. Although *LAR* is considered an important gene in the proanthocyanidin biosynthesis (Tanner *et al.*, 2003; Pang *et al.*, 2007; Liu *et al.*, 2016), many Brassicales species are well known for their ability to accumulate oligomeric proanthocyanidins. For example, the proanthocyanidin biosynthesis in *A. thaliana* was elucidated through a collection of transparent testa mutants that harbor a block in one of the proanthocyanin biosynthesis genes (Appelhaagen *et al.*, 2014).

Table 3: Annotation of the flavonoid biosynthesis genes in *T. majus*. Candidates lacking a residue in their predicted amino acid sequence that is considered crucial for their predicted function are marked with an asterisk.

Function	GeneID
<i>CHS</i>	TM29886*
<i>CHI</i>	TM11609
<i>F3H</i>	TM10740, TM15536
<i>F3'H</i>	TM28993, TM16564
<i>F3'5'H</i>	TM00413
<i>DFR</i>	TM17326, TM31282, TM31283
<i>ANS</i>	TM10587*
<i>arGST (TT19)</i>	TM05556
<i>A3GT</i>	TM04620

<i>FLS</i>	TM27596
<i>LAR</i>	-
<i>ANR</i>	TM02079*
Anthocyanin MYB	TM05502, TM05505
Flavonol MYB	TM09949
bHLH42/TT8	TM10100
TTG1	TM12689

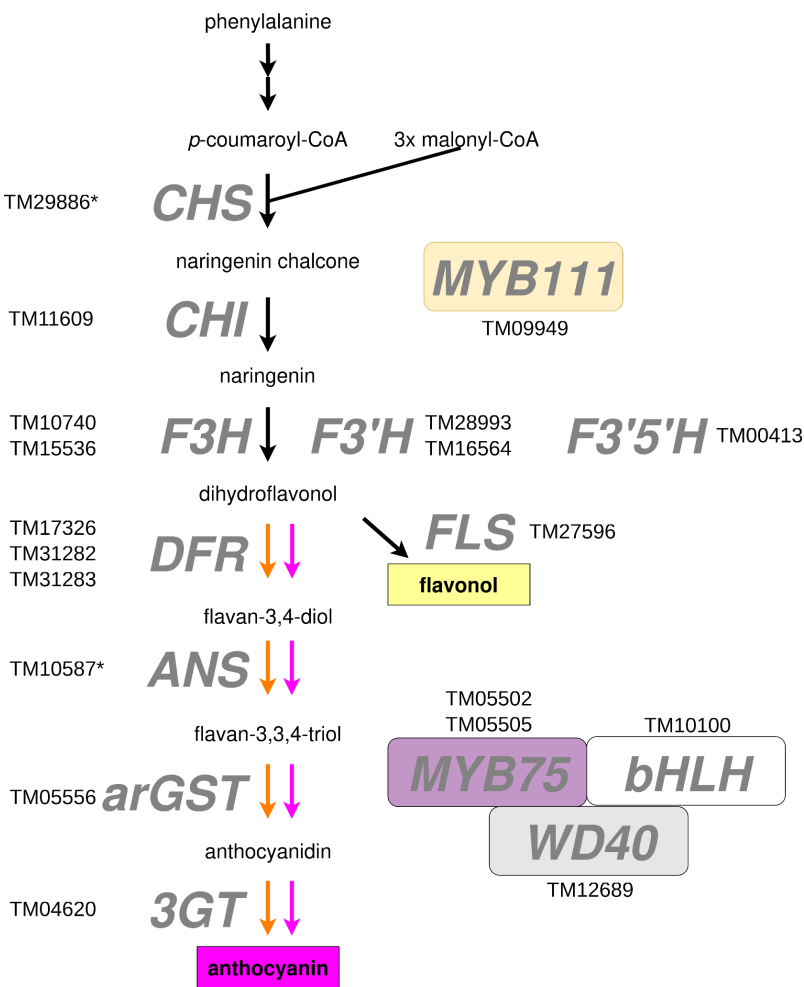


Fig. 1: Identified candidate genes in the flavonoid biosynthesis. Candidates lacking an amino acid residue in the derived amino acid sequence that is considered crucial for their predicted function are marked with an asterisk. The layout of this figure is based on (Horz *et al.*, 2024).

Activity of the flavonoid biosynthesis genes was assessed based on RNA-seq data (SRR25909845) belonging to seed samples (**Fig. 2**). Many genes of the anthocyanin biosynthesis are also part of the proanthocyanidin biosynthesis, which could be active in seeds of *T. majus*. Previous studies have reported important roles and substantial levels of proanthocyanidins in seeds of *A. thaliana* and *Vitis vinifera* (Nakamura *et al.*, 2003; Debeaujon *et al.*, 2003). It is also plausible that no flavonols are produced in the seeds which could explain the inactivity of *FLS* and its regulator.

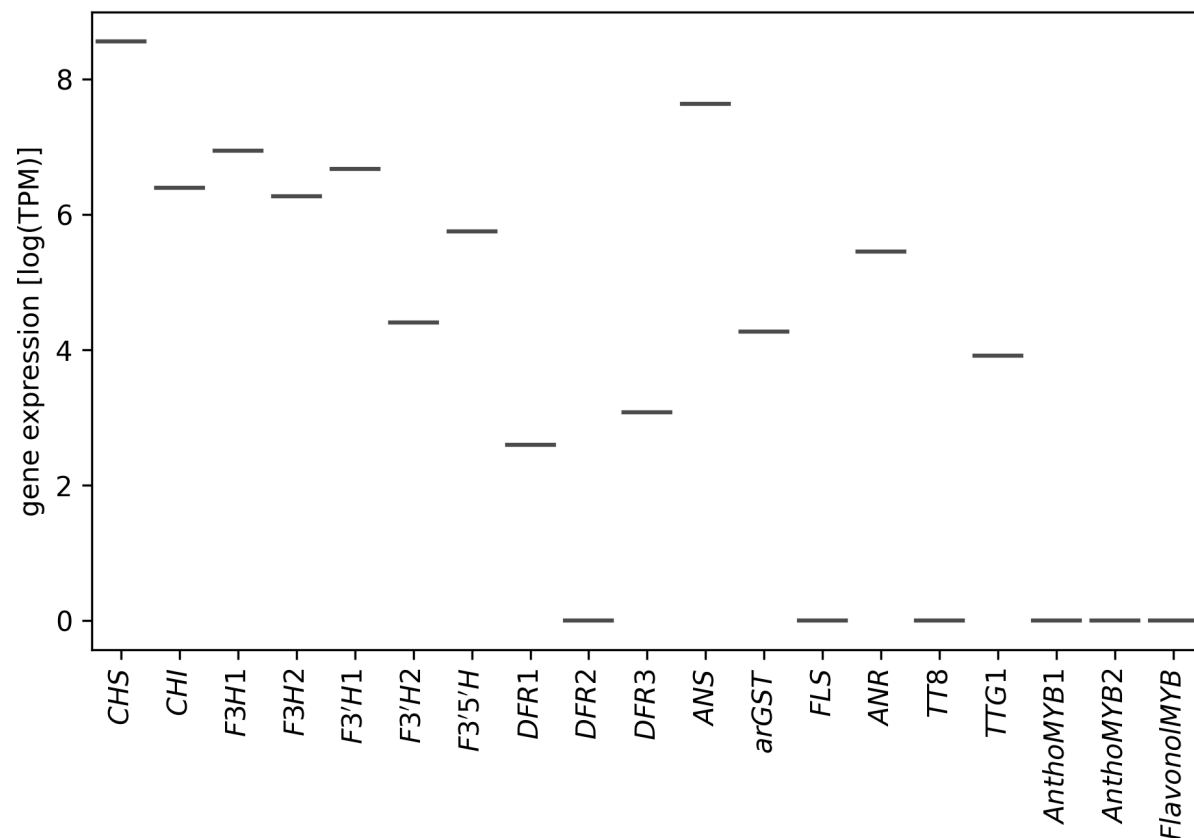


Fig. 2: Expression of flavonoid biosynthesis associated genes in seeds of *T. majus* (SRR25909845). Transcripts per million (TPMs) are displayed as log-transformed values.

Materials & methods

DNA extraction

T. majus L. (Accession TRO1, ID 47139) fruits were obtained from Genebank Gatersleben of the Leibniz Institute of Plant Genetics and Crop Plant Research, Germany. After soaking in water for 2 h, fruits were placed on Vermiculite and left in a controlled environment chamber at 22°C, 65 % humidity with a 10 h photoperiod (120 $\mu\text{mol m}^{-2} \text{s}^{-1}$ photosynthetically active

radiation) for a week. Seedlings were then transferred to potting substrate (germination potting compost (Compo, Münster, Germany) supplemented with 10 % (v/v) sand, 10 % (v/v) Perligran (Knauf Perlite, Germany), 2 g/l Triabon (Compo) and 1 g/l Sierrablen (Scotts, Heerlen, NL)) which had been steamed at 100°C for two hours and treated with Neudomück® (W. Neudorf GmbH KG, Emmerthal, Germany). Seventeen days after sawing, the photoperiod was changed to 16 h light. Plantlets were transferred to larger pots filled with steamed and Neudomück®-treated potting compost with 10 % (v/v) Perligran on day 32 after sawing. The largest leaves with parts of the petioles were harvested from a single plant at two time points (35 and 40 days after sawing), placed on ice and used immediately to isolate genomic DNA.

Leaf material was homogenized by grinding in liquid nitrogen. High molecular weight DNA was extracted using a modified CTAB-based protocol as previously described (Siadjeu *et al.*, 2020). DNA quality was assessed via NanoDrop measurement, agarose gel electrophoresis, and Qubit measurement as previously described (Horz *et al.*, 2024). Short DNA fragments were depleted with the Short Read Eliminator kit (Pacific Biosciences) following the suppliers instructions.

ONT sequencing

Nanopore sequencing libraries were prepared with the SQK-LSK109 ligation-based kit (Oxford Nanopore Technologies) using 1 µg of DNA per library and following the suppliers' instructions. ONT sequencing was performed on a MinION-Mk1B with R9.4.1 flow cells. Flow cell washing was performed upon blockage of a large proportion of nanopores followed by loading of a fresh library to achieve optimal performance of the flow cells. Basecalling of the raw sequencing data was performed with Guppy v6.4.6 (ONT) in high accuracy on a GPU in the de.NBI cloud. The customized Python script FASTQ_stats.py was applied to calculate statistics of the obtained sequencing data (<https://github.com/bpucker/GenomeAssembly>).

Genome sequence assembly

Initial assemblies were generated with Shasta v0.11.1 (Shafin *et al.*, 2020) to assess coverage of the sequencing data. The final assembly was generated based on all reads and NextDenovo v2.5.2 (Hu *et al.*, 2024) with the parameters genome_size = 800m, read_cutoff = 10k, seed_depth=45. The assembly statistics were compared with contig_stats3.py (Meckoni *et al.*, 2023). BUSCO v3.0.2 (Simão *et al.*, 2015; Manni *et al.*, 2021) was run in the genome mode to assess the completeness of both assemblies. The assembly generated by NextDenovo was selected as the representative genome sequence of *Tropaeolum majus*. The assembly was processed with clean_genomie_FASTA.py to reduce the sequence names to short identifiers without any special characters. Next, the genome sequence was screened for bacterial/fungal contamination contigs using assembly_wb_screen.py (Meckoni *et al.*, 2023).

Structural and functional annotation

Different approaches were conducted to predict protein encoding genes based on the assembled genome sequence. All results were compared with BUSCO v3.0.2 (Simão *et al.*, 2015; Manni *et al.*, 2021) run in the protein mode with embryophyta_odb10 as reference. RNA-seq reads of *T. majus* (SRR25909845) were retrieved from the Sequence Read Archive (Katz *et al.*, 2022) and aligned to the assembled genome sequence with STAR v2.7.11b (Dobin *et al.*, 2013; Dobin & Gingeras, 2015). BRAKER v3 (Gabriel *et al.*, 2023) was run with the aligned RNA-seq reads as hints. Another BRAKER run was supplied with protein hints derived from Viridiplantae (NCBI, 2023-05-01). AUGUSTUS v3.3 (Stanke *et al.*, 2006) was run in *ab initio* mode using previously optimized parameters for the detection of non-canonical splice sites (Pucker *et al.*, 2017). GeMoMa v1.9 (Keilwagen *et al.*, 2016, 2019) was run with protein hints derived from the following species *Carica papaya* (GCF_000150535.2) (Ming *et al.*, 2008), *Theobroma cacao* (GCF_000208745.1) (Argout *et al.*, 2011), *Arabidopsis halleri* (GCA_964271285.1), *A. thaliana* (GCF_000001735.4) (The Arabidopsis Genome Initiative, 2000; Sloan *et al.*, 2018), *Eutrema salsugineum* (GCF_000478725.1) (Yang *et al.*, 2013), and *Thlaspi arvense* (GCA_911865555.2) (Nunn *et al.*, 2022). The structural annotation produced by GeMoMa based on the comprehensive set of protein hints was selected as representative annotation. A corresponding functional annotation was generated with construct_anno.py (Pucker & Iorizzo, 2023) based on knowledge about gene functions in *A. thaliana* (Lamesch *et al.*, 2012; Cheng *et al.*, 2017).

Gene expression analysis

Kallisto v044 (Bray *et al.*, 2016) was used to process the publicly available RNA-seq data. Customized Python scripts were applied for the generation of a final count table (Pucker & Iorizzo, 2023). A figure summarizing the expression of flavonoid biosynthesis genes was generated based on normalized expression data (transcripts per million, TPMs) with a Python script using the matplotlib, pandas, scipy, numpy, and seaborn modules (Hunter, 2007; Virtanen *et al.*, 2020; Harris *et al.*, 2020; Waskom, 2021; Pucker & Iorizzo, 2023; The pandas development team, 2024).

Data availability

All data sets underlying this study are publicly available. Sequencing data have been deposited at the European Nucleotide Archive (PRJEB68329). The assembled genome sequence and corresponding annotation have been published via LeoPARD (https://leopard.tu-braunschweig.de/receive/dbbs_mods_00078092).

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 thank GeneBank at the IPK Gatersleben for providing *T. majus* fruits, Anita Backenköhler (Institute of Pharmaceutical Biology) for raising plants and all members of the research group Plant Biotechnology and Bioinformatics for discussion and support.

Author contributions

BP and UW conceived the project. RF extracted DNA and performed the sequencing. RF and BP conducted the bioinformatic analyses. UW and BP wrote the manuscript. RF, UW, and BP reviewed the final version of the manuscript and consented to its submission.

References

- Acurio L, Salazar D, Castillo B, Santiana C, Martínez-Monzó J, Igual M. 2024.** Characterization of Second-Generation Snacks Manufactured from Andean Tubers and Tuberous Root Flours. *Foods* **13**: 51. doi: 10.3390/foods13010051.
- Albrecht U, Stefenelli U, Stange R. 2023.** A combination of *Tropaeolum majus* herb and *Armoracia rusticana* root for the treatment of acute bronchitis. *Phytomedicine* **116**: 154838. doi: 10.1016/j.phymed.2023.154838.
- Andersson L, Andersson S. 2000.** A molecular phylogeny of Tropaeolaceae and its systematic implications. *TAXON* **49**: 721–736. doi: 10.2307/1223973.
- Appelhaugen I, Thiedig K, Nordholt N, Schmidt N, Huep G, Sagasser M, Weisshaar B. 2014.** Update on transparent testa mutants from Arabidopsis thaliana: characterisation of new alleles from an isogenic collection. *Planta* **240**: 955–970. doi: 10.1007/s00425-014-2088-0.
- Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, et al. 2011.** The genome of Theobroma cacao. *Nature Genetics* **43**: 101–108. doi: 10.1038/ng.736.
- Bray NL, Pimentel H, Melsted P, Pachter L. 2016.** Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**: 525–527. doi: 10.1038/nbt.3519.
- Cardinal-McTeague WM, Sytsma KJ, Hall JC. 2016.** Biogeography and diversification of Brassicales: A 103 million year tale. *Molecular Phylogenetics and Evolution* **99**: 204–224. doi: 10.1016/j.ympev.2016.02.021.
- Česlová L, Klikarová J, Šalomounová T. 2023.** The content and profile of biologically active compounds present in individual parts of nasturtium (*Tropaeolum majus* L.): comprehensive study. *European Food Research and Technology* **249**: 413–428. doi: 10.1007/s00217-022-04126-4.
- Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. 2017.** AraPort11: a complete reannotation of the Arabidopsis thaliana reference genome. *The Plant Journal* **89**: 789–804. doi: 10.1111/tpj.13415.
- Debeaujon I, Nesi N, Perez P, Devic M, Grandjean O, Caboche M, Lepiniec L. 2003.** Proanthocyanidin-Accumulating Cells in Arabidopsis Testa: Regulation of Differentiation and Role in Seed Development. *The Plant Cell* **15**: 2514–2531. doi: 10.1105/tpc.014043.
- 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. doi: 10.1093/bioinformatics/bts635.
- Dobin A, Gingeras TR. 2015.** Mapping RNA-seq Reads with STAR. *Current Protocols in Bioinformatics* **51**: 11.14.1–11.14.19. doi: 10.1002/0471250953.bi1114s51.
- 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. doi: 10.1101/2023.06.10.544449.

Garzón GA, Manns DC, Riedl K, Schwartz SJ, Padilla-Zakour O. 2015. Identification of phenolic compounds in petals of nasturtium flowers (*Tropaeolum majus*) by high-performance liquid chromatography coupled to mass spectrometry and determination of oxygen radical absorbance capacity (ORAC). *Journal of Agricultural and Food Chemistry* **63**: 1803–1811. doi: 10.1021/jf503366c.

Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. 2020. Array programming with NumPy. *Nature* **585**: 357–362. doi: 10.1038/s41586-020-2649-2.

Horz JM, Wolff K, Friedhoff R, Pucker B. 2024. Genome sequence of the ornamental plant *Digitalis purpurea* reveals the molecular basis of flower color and morphology variation. : 2024.02.14.580303. doi: 10.1101/2024.02.14.580303.

Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, Li J, Sandoval JR, Cooper DN, Ye K, et al. 2024. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biology* **25**: 107. doi: 10.1186/s13059-024-03252-4.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**: 90–95. doi: 10.1109/MCSE.2007.55.

Jensen JK, Schultink A, Keegstra K, Wilkerson CG, Pauly M. 2012. RNA-Seq Analysis of Developing Nasturtium Seeds (*Tropaeolum majus*): Identification and Characterization of an Additional Galactosyltransferase Involved in Xyloglucan Biosynthesis. *Molecular Plant* **5**: 984–992. doi: 10.1093/mp/sss032.

Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O’Sullivan C. 2022. The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Research* **50**: D387–D390. doi: 10.1093/nar/gkab1053.

Keilwagen J, Hartung F, Grau J. 2019. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. *Methods in Molecular Biology (Clifton, N.J.)* **1962**: 161–177. doi: 10.1007/978-1-4939-9173-0_9.

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

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* **40**: D1202–D1210. doi: 10.1093/nar/gkr1090.

Liu C, Wang X, Shulaev V, Dixon RA. 2016. A role for leucoanthocyanidin reductase in the extension of proanthocyanidins. *Nature Plants* **2**: 1–7. doi: 10.1038/nplants.2016.182.

Lu N. 2024. Revisiting decade-old questions in proanthocyanidin biosynthesis: current understanding and new challenges. *Frontiers in Plant Science* **15**. doi: 10.3389/fpls.2024.1373975.

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. doi: 10.1093/molbev/msab199.

Marks RA, Hotelling S, Frandsen PB, VanBuren R. 2021. Representation and participation across 20 years of plant genome sequencing. *Nature Plants* **7**: 1571–1578. doi: 10.1038/s41477-021-01031-8.

Meckoni SN, Nass B, Pucker B. 2023. Phylogenetic placement of *Ceratophyllum submersum* based on a complete plastome sequence derived from nanopore long read sequencing data. *BMC Research Notes* **16**: 187. doi: 10.1186/s13104-023-06459-z.

- Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KLT, et al. 2008.** The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**: 991–996. doi: 10.1038/nature06856.
- Missouri Botanical Garden. 2024.** Tropaeolaceae.
- Nakamura Y, Tsuji S, Tonogai Y. 2003.** Analysis of Proanthocyanidins in Grape Seed Extracts, Health Foods and Grape Seed Oils. *Journal of Health Science* **49**: 45–54. doi: 10.1248/jhs.49.45.
- Nesi N, Lucas M-O, Auger B, Baron C, Lécureuil A, Guerche P, Kronenberger J, Lepiniec L, Debeaujon I, Renard M. 2009.** The promoter of the *Arabidopsis thaliana* BAN gene is active in proanthocyanidin-accumulating cells of the *Brassica napus* seed coat. *Plant Cell Reports* **28**: 601–617. doi: 10.1007/s00299-008-0667-x.
- Nunn A, Rodríguez-Arévalo I, Tandukar Z, Frels K, Contreras-Garrido A, Carbonell-Bejerano P, Zhang P, Ramos Cruz D, Jandrasits K, Lanz C, et al. 2022.** Chromosome-level *Thlaspi arvense* genome provides new tools for translational research and for a newly domesticated cash cover crop of the cooler climates. *Plant Biotechnology Journal* **20**: 944–963. doi: 10.1111/pbi.13775.
- Pang Y, Peel GJ, Wright E, Wang Z, Dixon RA. 2007.** Early steps in proanthocyanidin biosynthesis in the model legume *Medicago truncatula*. *Plant Physiology* **145**: 601–615. doi: 10.1104/pp.107.107326.
- Pucker B, Brockington SF. 2018.** Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. *BMC Genomics* **19**: 980. doi: 10.1186/s12864-018-5360-z.
- Pucker B, Friedhoff R, Wittstock U. 2024.** Genome sequence and annotation of the medicinal plant *Tropaeolum majus*. doi: 10.24355/dbbs.084-202410230959-0.
- 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. doi: 10.1186/s13104-017-2985-y.
- Pucker B, Iorizzo M. 2023.** Apiaceae FNS I originated from F3H through tandem gene duplication. *PLOS ONE* **18**: e0280155. doi: 10.1371/journal.pone.0280155.
- 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. doi: 10.3390/plants9091103.
- 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. doi: 10.1038/s41587-020-0503-6.
- 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. doi: 10.3390/genes11030274.
- 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. doi: 10.1093/bioinformatics/btv351.
- Sloan DB, Wu Z, Sharbrough J. 2018.** Correction of Persistent Errors in *Arabidopsis* Reference Mitochondrial Genomes. *The Plant Cell* **30**: 525–527. doi: 10.1105/tpc.18.00024.
- 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. doi: 10.1093/nar/gkl200.
- Stracke R, Ishihara H, Huep G, Barsch A, Mehrrens F, Niehaus K, Weisshaar B. 2007.** Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the *Arabidopsis thaliana* seedling. *The Plant Journal: For Cell and Molecular Biology* **50**: 660–677. doi: 10.1111/j.1365-3113X.2007.03078.x.

- Stracke R, Jahns O, Keck M, Tohge T, Niehaus K, Fernie AR, Weisshaar B. 2010.** Analysis of PRODUCTION OF FLAVONOL GLYCOSIDES-dependent flavonol glycoside accumulation in *Arabidopsis thaliana* plants reveals MYB11-, MYB12- and MYB111-independent flavonol glycoside accumulation. *New Phytologist* **188**: 985–1000. doi: 10.1111/j.1469-8137.2010.03421.x.
- Tanner GJ, Francki KT, Abrahams S, Watson JM, Larkin PJ, Ashton AR. 2003.** Proanthocyanidin Biosynthesis in Plants: PURIFICATION OF LEGUME LEUCOANTHOCYANIDIN REDUCTASE AND MOLECULAR CLONING OF ITS cDNA *. *Journal of Biological Chemistry* **278**: 31647–31656. doi: 10.1074/jbc.M302783200.
- The Arabidopsis Genome Initiative. 2000.** Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815. doi: 10.1038/35048692.
- The pandas development team. 2024.** Pandas. doi: 10.5281/zenodo.13819579.
- Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. 2020.** SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**: 261–272. doi: 10.1038/s41592-019-0686-2.
- Waskom M. 2021.** seaborn: statistical data visualization. *Journal of Open Source Software* **6**: 3021. doi: 10.21105/joss.03021.
- Winkel-Shirley B. 2001.** Flavonoid Biosynthesis. A Colorful Model for Genetics, Biochemistry, Cell Biology, and Biotechnology. *Plant Physiology* **126**: 485–493. doi: 10.1104/pp.126.2.485.
- Yang R, Jarvis DE, Chen H, Beilstein MA, Grimwood J, Jenkins J, Shu S, Prochnik S, Xin M, Ma C, et al. 2013.** The Reference Genome of the Halophytic Plant *Eutrema salsugineum*. *Frontiers in Plant Science* **4**: 46. doi: 10.3389/fpls.2013.00046.