

1 **A highly contiguous genome assembly of *Cyclamen persicum* to accelerate functional**
2 **genomics and breeding**

3 Running title: Genome assembly of *Cyclamen persicum*

4

5 Kenta Shirasawa^{1*}, Yusuke Akita², Yuki Mizunoe³, Takejiro Takamura⁴

6

7 ¹ Department of Frontier Research and Development, Kazusa DNA Research Institute, Kisarazu,
8 Chiba 292-0818, Japan

9 ² Department of Life Science and Green Chemistry, Faculty of Engineering, Saitama Institute of
10 Technology, Fukaya, Saitama 369-0293, Japan

11 ³ Faculty of Agriculture, Kyushu University, Fukuoka, Fukuoka 819-0395, Japan

12 ⁴ Faculty of Agriculture, Kagawa University, Miki, Kagawa 761-0795, Japan

13

14 *To whom correspondence should be addressed: Tel. +81 438 52 3935; Email
15 shirasaw@kazusa.or.jp

16

1 **Abstract**

2 *Cyclamen* is an economically important ornamental plant widely cultivated for its diverse floral
3 characteristics and adaptation to cool climates. Despite its horticultural significance, genomic
4 resources for this species remain limited, hindering molecular studies and genomics-assisted
5 breeding. Here, we report the first highly contiguous nuclear genome assembly of *C. persicum*
6 generated using high-fidelity long-read sequencing. The assembled genome spans 1.48 Gb,
7 consisting of 126 contigs with an N50 length of 52.3 Mb. Telomeric repeat analysis identified
8 eight contigs containing telomeric sequences at both ends, suggesting the presence of
9 near-complete chromosome assemblies. Genome completeness assessment using BUSCO
10 indicated 98.1% completeness. Repetitive sequences occupied 82.9% of the assembly, with long
11 terminal repeat retrotransposons accounting for 42.1% of the genome. A total of 40,223
12 protein-coding genes were predicted, with a complete BUSCO score of 95.7%. Comparative
13 orthogroup analysis with five representative eudicot species identified 430 orthogroups specific
14 to *C. persicum* and 363 orthogroups shared exclusively between *C. persicum* and *Primula*
15 *kwangtungensis*, indicating the presence of both lineage-specific and Primulaceae-conserved
16 gene families. These findings provide critical insights into gene family evolution within
17 Primulaceae and establish an essential comparative framework for future genomic studies. The
18 genome resource presented here provides an invaluable foundation for investigating genome
19 evolution, gene function, and trait-associated loci in cyclamen, effectively facilitating molecular
20 breeding and genetic improvement in this ornamental species.

21 **Keywords:** gene annotation; molecular breeding; ornamental plant; Primulaceae; repetitive
22 elements

23

1 **1. Introduction**

2 Understanding plant genome structure, evolution, and functional diversity is a central goal of
 3 plant genomics. Over the past two decades, advances in high-throughput sequencing
 4 technologies and bioinformatic methodologies have enabled the generation of reference
 5 genomes for a wide range of angiosperms^{1,2}, thereby accelerating comparative genomics and
 6 functional studies³. However, many horticulturally important ornamental plants remain
 7 underrepresented in genomic databases. This is often due to their complex genomic features,
 8 such as large genome sizes and high proportions of repetitive elements, thus limiting a
 9 molecular-level understanding of the traits that underpin their economic and biological value.

10 The genus *Cyclamen* (Primulaceae) comprises approximately 25 species distributed
 11 across Europe, North Africa, and the Middle East (<https://powo.science.kew.org>). Species of
 12 *Cyclamen* are widely cultivated as ornamental pot and garden plants and are characterized by
 13 diverse floral morphologies, vivid flower colors, tuberous growth habits, and adaptation to cool
 14 climates^{4,5}. Although the representative diploid chromosome number in the genus is $2n = 48$,
 15 *Cyclamen* species exhibit substantial variation in genome size, chromosome number, and
 16 ecological niche, making the genus an intriguing system for studying genome evolution and
 17 speciation within the Primulaceae⁵⁻⁸. However, the genomic mechanisms driving this structural
 18 and size variation, as well as the dynamics of repetitive elements, remain largely unresolved due
 19 to the lack of high-quality reference genomes. Among the genus members, *Cyclamen persicum*
 20 is of particular importance, as nearly all modern cultivated cyclamen varieties have been
 21 developed from this species. Cultivars of *C. persicum* display extensive variation in flower
 22 color and morphology, making the species a valuable system for studying the genetic basis of
 23 ornamental traits. Thus, the establishment of a high-quality reference genome for *C. persicum*
 24 will not only advance our understanding of genome evolution within *Cyclamen* but also provide

1 a cornerstone for future pan-genomic and evolutionary studies across the genus.

2 Molecular biological studies in *Cyclamen* have increased in recent years. Several
3 studies have identified genes involved in flower color, floral organ development, and related
4 traits^{9–18}. Additionally, molecular resources have been generated through chloroplast genome
5 sequencing and RNA sequencing–based transcriptome analyses^{19–22}. While these studies have
6 provided useful insights, they have been conducted largely at the single-gene or transcriptome
7 level, without a structural genomic context. Consequently, they offer limited information on
8 genome organization, repetitive element dynamics, and the evolutionary context of gene
9 families. Moreover, the lack of a reference nuclear genome has hindered comprehensive
10 genomic analyses and the high-throughput development of molecular markers in *Cyclamen*. A
11 highly contiguous genome assembly of *C. persicum* will provide a framework for systematic
12 gene annotation, investigation of genome structure, and comparative analyses within the
13 Primulaceae and related taxa.

14 Plant genome assembly, however, poses significant challenges, particularly in
15 perennial ornamental species characterized by large genome sizes, high heterozygosity, and
16 complex genome architectures shaped by repetitive elements and structural variation. In recent
17 years, whole-genome sequencing has been increasingly applied to horticultural plants with such
18 complex genomes, driven by rapid advances in long-read sequencing technologies with high
19 base accuracy, such as PacBio HiFi^{23,24}. These advances now enable the generation of
20 near-complete, highly contiguous assemblies even for species with intricate genomic
21 structures^{25,26}. Furthermore, the emergence of deep-learning-based gene prediction workflows
22 has facilitated highly accurate, *ab initio* gene prediction even in the absence of comprehensive
23 transcriptomic resources²⁷.

24 In this study, we report a highly contiguous reference genome assembly of *C.*

persicum generated using PacBio HiFi long-read sequencing. This high-quality genome assembly represents the first comprehensive nuclear genomic resource for the genus *Cyclamen*, providing a foundational platform to address evolutionary questions within the Ericales and establishing a robust molecular resource to accelerate breeding and functional genomics in this highly valued ornamental crop.

2. Materials and methods

2.1. Plant materials

Young leaves of *C. persicum* ‘KN purple’, cultivated in a greenhouse at Faculty of Agriculture, Kagawa University, were collected for genome DNA extraction. The ‘KN purple’ plants had been propagated by repeated inbreeding including self-pollination.

2.2. Genome size estimation

To construct a short-read library, a PCR-free library was prepared using the Swift 2S Turbo Flexible DNA Library Kit (Swift Biosciences, Ann Arbor, MI, USA). This library was then converted into a DNA nanoball sequencing library using the MGI Easy Universal Library Conversion Kit (MGI Tech, Shenzhen, China) and sequenced on a DNBSEQ G400 instrument (MGI Tech) to generate 2×100 -bp paired-end reads. The genome size of *C. persicum* was estimated by *k*-mer frequency analysis ($k = 17$) using Jellyfish²⁸ v2.3.0 (*k*-mer size = 17).

2.3. Genome assembly

Genomic DNA was extracted from young leaves using the Genomic Tip (Qiagen, Hilden, Germany). The genomic DNA was sheared to an average fragment size of 30 kb using Megaruptor 2 (Diagenode, Liege, Belgium) in the Large Fragment Hydropore mode. The

sheared DNA was then used for HiFi SMRTbell library preparation using the SMRTbell Express Template Prep Kit 3.0 (PacBio, Menlo Park, CA, USA). The resultant library was separated on BluePippin (Sage Science, Beverly, CA, USA) to remove short DNA fragments (<15 kb) and sequenced using SMRT Cell 8 M on the Sequel II system (PacBio).

The obtained HiFi reads were assembled using Hifiasm²⁹ v0.16.1 with default parameters. Organelle genome sequences, identified by sequence similarity searches of the reported plastid and mitochondrial genome sequences of cyclamen² and primula³⁰, respectively, using Minimap2³¹ v2.24, were eliminated. Assembly completeness was assessed with embryophyta_odb10 lineage data using Benchmarking Universal Single-Copy Orthologs (BUSCO)³² v5.2.2. Telomere sequences containing repeats of a 7-bp motif (5'-TTTAGGG-3') were searched using the search subcommand of Tidk³³ v0.2.1 with a window size of 1000 bp.

2.4. Gene and repeat prediction

Protein-coding genes were predicted using the *ab initio* strategy of Helixer²⁷. The completeness of the predicted gene models was evaluated with the embryophyta_odb10 data using BUSCO³² v5.2.2. Repetitive sequences in the assembly were identified using RepeatMasker v4.1.6 (<https://www.repeatmasker.org>), based on repeat sequences registered in Repbase and a *de novo* repeat library built with RepeatModeler v2.0.4 (<https://www.repeatmasker.org>).

2.5. Comparative genomics analysis

Comparative orthogroup analysis was conducted using OrthoFinder v2.5³⁴ with predicted protein sequences from *C. persicum* and five representative eudicot species, including *Arabidopsis thaliana* (Araport11, <https://www.arabidopsis.org>)³⁵, *Vitis vinifera* (12X, <https://www.genoscope.cns.fr/vitis>)³⁶, *Camellia sinensis* (CsSME v1.0, GenBank accession

1 number of GCA_041154535.1)³⁷, *Primula kwangtungensis* (Prkwt_v. 1.0,
2 GCA_054094925.1)³⁸, and *Rhododendron ripense* (RRL_r1.1, GCA_019656295.1)³.
3 Orthologous gene groups shared among species and species-specific orthogroups were
4 visualized using an UpSet plot³⁹.

5

6 **3. Results**

7 **3.1. Genome assembly and annotation**

8 To estimate the *C. persicum* genome size, clean short-read data (70.4 Gb) were subjected to
9 *k*-mer distribution analysis. The results indicated that the *C. persicum* ‘KN purple’ genome
10 exhibited low heterozygosity, with an estimated haploid genome size of 1.8 Gb (Figure 1). A
11 total of 2.9 million HiFi reads (41.3 Gb; 22.9× coverage of the estimated genome size), were
12 obtained from three SMRT cells and assembled into 798 contigs. A total of 672 potential
13 contaminant sequences (32.4 Mb in total) from organelle genomes were removed. Consequently,
14 126 contigs totaling 1.48 Gb, with an N50 of 52.3 Mb, were obtained (Table 1). These 126
15 contigs represented the genome assembly of *C. persicum* and were designated CPE_r1.0.
16 Notably, more than 95% of the assembled sequence (1.41 Gb) was represented by 40 contigs
17 (Figure 2). Telomere repeats were found at both ends of 8 contigs and at one end of 26 contigs
18 (Figure 2), suggesting that these 8 contigs were assembled at the complete telomere-to-telomere
19 chromosome level. A total of 40,223 protein-coding genes were predicted from the genomic
20 sequence (Table 1). Repetitive sequences occupied 1.23 Gb (82.9%) of the genome, of which
21 LTR retroelements (624.1 Mb; 42.1%) represented the dominant repeat type (Table 2). The
22 accuracy of the genome assembly and gene predictions were supported by BUSCO scores of
23 98.1% and 95.7%, respectively (Table 3).

24

1 3.2. Comparative orthogroup analysis

2 To investigate gene family conservation and diversification in *C. persicum*, comparative
3 orthogroup analysis was conducted using the predicted protein sequences from *C. persicum* and
4 five representative eudicot species: *Arabidopsis thaliana*, *Vitis vinifera*, *Camellia sinensis*,
5 *Primula kwangtungensis*, and *Rhododendron ripense* (Figure 3).

6 Orthogroup analysis revealed extensive conservation of gene families among the
7 examined species, with numerous orthogroups shared across multiple eudicot lineages. Notably,
8 430 orthogroups were identified exclusively in *C. persicum*, indicating the presence of
9 lineage-specific gene families that may have arisen during the evolution of *Cyclamen*. In
10 addition, 363 orthogroups were shared exclusively between *C. persicum* and *P. kwangtungensis*,
11 the two representatives of Primulaceae included in this study, suggesting the retention of gene
12 families that are characteristic of the family. The analysis also identified 474 orthogroups shared
13 among all examined species except *A. thaliana*, and 1,047 orthogroups shared among all species
14 except *R. ripense*. These patterns may reflect lineage-specific gene loss, divergence, or
15 differences in genome annotation among species. Overall, the results indicate that the *C.*
16 *persicum* genome contains both highly conserved eudicot gene families and lineage-specific
17 gene repertoires, providing a useful framework for future comparative and evolutionary studies
18 within Primulaceae and Ericales.

19

20 4. Discussion

21 In this study, we generated a highly contiguous reference genome assembly of *C. persicum*
22 (CPE_r1.0), resolving the 1.48 Gb genome into just 126 contigs with a contig N50 length of
23 52.3 Mb (Table 1). The discrepancy between the estimated genome size (Figure 1) and assembly
24 length is likely attributable to collapsed highly repetitive regions (Table 2) and/or

1 overestimation in *k*-mer analysis⁴⁰. The high BUSCO completeness score (98.1%) confirms that
 2 most of the conserved gene space was successfully captured (Table 3). Although *C. persicum*
 3 generally exhibits strong inbreeding depression⁴¹ and high heterozygosity in cultivated lines,
 4 our *k*-mer analysis revealed relatively low heterozygosity in the targeted accession (Figure 1).
 5 This indicates that the strategic selection of a partially inbred line successfully resolved the
 6 computational haplotype-purging challenges typical of outcrossing species, demonstrating the
 7 effectiveness of high-accuracy PacBio HiFi long-read sequencing for complex plant genomes.

8 Repetitive sequences occupied 82.9% of the CPE_r1.0 assembly, dominated by LTR
 9 retrotransposons (Table 2). Notably, 35.1% of these repeats remained unclassified, suggesting a
 10 substantial reservoir of lineage-specific elements that likely drove genomic expansion in
 11 *Cyclamen*, consistent with other large-genome angiosperms⁴². Despite this heavy repeat burden,
 12 our annotation using the *ab initio* pipeline of Helixer successfully identified 40,223
 13 protein-coding genes (Table 1) with a high BUSCO completeness of 95.7% (Table 3). This level
 14 of completeness, achieved without transcriptome-assisted training, confirms that CPE_r1.0
 15 provides robust and reliable gene models suitable for downstream comparative and functional
 16 analyses, with gene numbers aligning well with those of related eudicots^{43,44}.

17 Our comparative orthogroup analysis highlighted 363 gene families shared exclusively
 18 between *C. persicum* and *P. kwangtungensis*, representing core family-specific synapomorphies
 19 conserved within the Primulaceae family (Figure 3). These gene families provide an important
 20 starting point for identifying traits conserved within Primulaceae. Furthermore, the 430 *C.*
 21 *persicum*-specific orthogroups represent prime candidates underlying genus-specific traits, such
 22 as unique floral ontogeny, tuberous growth habits, and specialized volatile profiles. CPE_r1.0
 23 now serves as a critical genomic resource to integrate fragmented transcriptomic data^{19–22} and
 24 leverage established transformation systems^{45,46} to facilitate transgenic phenotypic

modifications^{9,47}. From an applied perspective, CPE_r1.0 offers a valuable framework to overcome conventional breeding bottlenecks in this outcrossing species⁴⁸, thereby accelerating genomics-assisted breeding and functional genomics in this ornamental crop.

Acknowledgments

We thank Y. Kishida, C. Minami, K. Ozawa, H. Tsuruoka, and A. Watanabe (Kazusa DNA Research Institute) for technical assistance.

Funding

This work was supported by JSPS KAKENHI (JP22H05172, JP22H05181, and JP24K08899) and the Kazusa DNA Research Institute Foundation.

Conflict of interest

The authors declare that there is no conflict of interest.

Data availability

Raw sequence reads and assembled sequences are available at DDBJ (BioProject accession number PRJDB42699). The genome assembly and gene annotation files are available in Kazusa Genome Atlas (<https://genome.kazusa.or.jp>).

References

1. Kersey, P. J. 2019, Plant genome sequences: past, present, future. *Curr. Opin. Plant Biol.*, **48**, 1–8.
2. Sun, Y., Shang, L., Zhu, Q.-H., Fan, L., and Guo, L. 2022, Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci.*, **27**, 391–401.

- 1 3. Shirasawa, K., Kobayashi, N., Nakatsuka, A., Ohta, H., and Isobe, S. 2021, Whole-genome
2 sequencing and analysis of two azaleas, *Rhododendron ripense* and *Rhododendron*
3 *kiyosumense*. *DNA Res.*, **28**.
- 4 4. Anderberg, A. A., Trift, I., and Kallersjö, M. 2000, Phylogeny of *Cyclamen* L.
5 (Primulaceae): Evidence from morphology and sequence data from the internal transcribed
6 spacers of nuclear ribosomal DNA. *Plant Syst. Evol.*, **220**, 147–60.
- 7 5. Temsch, E. M., Temsch, W., Ehrendorfer-Schratt, L., and Greilhuber, J. 2010, Heavy metal
8 pollution, selection, and genome size: The species of the Žerjav study revisited with flow
9 cytometry. *J. Bot.*, **2010**, 1–11.
- 10 6. Bennett, S. T., and Grimshaw, J. M. 1991, Cytological studies in *Cyclamen* subg.
11 *Cyclamen* (Primulaceae). *Plant Syst. Evol.*, **176**, 135–43.
- 12 7. Pavlíček, T., Bureš, P., Horová, L., Raskina, O., and Nevo, E. 2008, Genome size
13 microscale divergence of *Cyclamen persicum* in Evolution Canyon, Israel. *Open Life Sci.*,
14 **3**, 83–90.
- 15 8. Yesson, C., and Culham, A. 2006, A phyloclimatic study of *Cyclamen*. *BMC Evol. Biol.*, **6**,
16 72.
- 17 9. Boase, M. R., Lewis, D. H., Davies, K. M., et al. 2010, Isolation and antisense suppression
18 of flavonoid 3', 5'-hydroxylase modifies flower pigments and colour in cyclamen. *BMC*
19 *Plant Biol.*, **10**, 107.
- 20 10. Akita, Y., Ishizaka, H., Nakayama, M., et al. 2010, Comparative analysis of floral
21 pigmentation between wild-type and white-flowered varieties of *Cyclamen graecum*. *J.*
22 *Hortic. Sci. Biotechnol.*, **85**, 437–43.
- 23 11. Tanaka, Y., Yamamura, T., and Terakawa, T. 2011, Identification and expression analysis of
24 the *Cyclamen persicum* MADS-box gene family. *Plant Biotechnol. (Tsukuba)*, **28**, 167–72.
- 25 12. Togami, J., Okuhara, H., Nakamura, N., et al. 2011, Isolation of cDNAs encoding
26 tetrahydroxychalcone 2'-glucosyltransferase activity from carnation, cyclamen, and
27 catharanthus. *Plant Biotechnol. (Tsukuba)*, **28**, 231–8.
- 28 13. Akita, Y., Kitamura, S., Hase, Y., et al. 2011, Isolation and characterization of the fragrant
29 cyclamen O-methyltransferase involved in flower coloration. *Planta*, **234**, 1127–36.
- 30 14. Kitamura, S., Akita, Y., Ishizaka, H., Narumi, I., and Tanaka, A. 2012, Molecular
31 characterization of an anthocyanin-related glutathione S-transferase gene in cyclamen. *J.*
32 *Plant Physiol.*, **169**, 636–42.
- 33 15. Mizunoe, Y., Kubota, S., Kanno, A., and Ozaki, Y. 2015, Morphological Variation and
34 AGAMOUS-like Gene Expression in Double Flowers of *Cyclamen persicum* Mill. *Hort. J.*,
35 **84**, 140–7.
- 36 16. Akita, Y., Kitamura, S., Mikami, R., and Ishizaka, H. 2017, Identification of functional

- 1 flavonol synthase genes from fragrant wild cyclamen (*Cyclamen purpurascens*). *J. Plant*
- 2 *Biochem. Biotechnol.*
- 3 17. Kang, X., Mikami, R., and Akita, Y. 2021, Characterization of 5-O-glucosyltransferase
- 4 involved in anthocyanin biosynthesis in *Cyclamen purpurascens*. *Plant Biotechnol.*
- 5 *(Tsukuba)*, **38**, 263–8.
- 6 18. Mikami, R., Kitamura, S., and Akita, Y. 2025, Identification of a functional *flavanone*
- 7 *3-hydroxylase* gene in a fragrant wild cyclamen species (*cyclamen purpurascens*). *Hort. J.*,
- 8 **94**, 364–73.
- 9 19. Xia, D., He, G., Wang, K., et al. 2022, Anthocyanins profiling analysis and RNA-Seq
- 10 revealed the dominating pigments and coloring mechanism in cyclamen flowers. *Biology*
- 11 *(Basel)*, **11**, 1721.
- 12 20. Mao, L., Zou, Q., Sun, Z., Dong, Q., and Cao, X. 2023, Insights into chloroplast genome
- 13 structure, intraspecific variation, and phylogeny of *Cyclamen* species (Myrsinoideae). *Sci.*
- 14 *Rep.*, **13**, 87.
- 15 21. Zhang, W., Ba, J., Guo, J., Jiang, Z., Zhang, Z., and Cui, Z. 2024, The first complete
- 16 chloroplast genome of *Cyclamen persicum* and its phylogenetic position within
- 17 Primulaceae. *Mitochondrial DNA B Resour.*, **9**, 939–42.
- 18 22. Hekmati, Z., Solouki, M., Emamjomeh, A., Zahiri, J., and Mirzaie-Asl, A. 2025,
- 19 Transcriptomic analysis of cyclamen persicum to identify invovled genes in triterpene
- 20 secondary metabolites pathway. *Biochem. Genet.*, **63**, 1509–26.
- 21 23. Harkess, A., Zhou, J., Xu, C., et al. 2017, The asparagus genome sheds light on the origin
- 22 and evolution of a young Y chromosome. *Nat. Commun.*, **8**, 1279.
- 23 24. Raymond, O., Gouzy, J., Just, J., et al. 2018, The Rosa genome provides new insights into
- 24 the domestication of modern roses. *Nat. Genet.*, **50**, 772–7.
- 25 25. Nashima, K., Shirasawa, K., Ghelfi, A., et al. 2021, Genome sequence of *Hydrangea*
- 26 *macrophylla* and its application in analysis of the double flower phenotype. *DNA Res.*, **28**.
- 27 26. Shirasawa, K., Ueta, S., Murakami, K., Abdelrahman, M., Kanno, A., and Isobe, S. 2022,
- 28 Chromosome-scale haplotype-phased genome assemblies of the male and female lines of
- 29 wild asparagus (*Asparagus kiusianus*), a dioecious plant species. *DNA Research*, **29**,
- 30 dsac002.
- 31 27. Stiehler, F., Steinborn, M., Scholz, S., Dey, D., Weber, A. P. M., and Denton, A. K. 2021,
- 32 Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning.
- 33 *Bioinformatics*, **36**, 5291–8.
- 34 28. Marçais, G., and Kingsford, C. 2011, A fast, lock-free approach for efficient parallel
- 35 counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–70.
- 36 29. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. 2021, Haplotype-resolved

- 1 de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods*, **18**, 170–5.
- 2 30. Wei, L., Liu, T.-J., Hao, G., Ge, X.-J., and Yan, H.-F. 2022, Comparative analyses of three
3 complete *Primula* mitogenomes with insights into mitogenome size variation in Ericales.
4 *BMC Genomics*, **23**, 770.
- 5 31. Li, H. 2018, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**,
6 3094–100.
- 7 32. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M.
8 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy
9 orthologs. *Bioinformatics*, **31**, 3210–2.
- 10 33. Brown, M. R., Manuel Gonzalez de La Rosa, P., and Blaxter, M. 2025, Tidk: A toolkit to
11 rapidly identify telomeric repeats from genomic datasets. *Bioinformatics*, **41**.
- 12 34. Emms, D. M., and Kelly, S. 2019, OrthoFinder: phylogenetic orthology inference for
13 comparative genomics. *Genome Biol.*, **20**, 238.
- 14 35. Cheng, C.-Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., and Town, C.
15 D. 2017, Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome.
16 *Plant J.*, **89**, 789–804.
- 17 36. Jaillon, O., Aury, J.-M., Noel, B., et al. 2007, The grapevine genome sequence suggests
18 ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–7.
- 19 37. Kawahara, Y., Tanaka, J., Takayama, K., et al. 2024, Chromosome-scale genome assembly
20 and characterization of top-quality Japanese green tea cultivar “Seimei.” *Plant Cell*
21 *Physiol.*, **65**, 1271–84.
- 22 38. Liu, T., Xiao, Y., Wu, X., et al. 2026, Chromosome-level genome assembly and annotation
23 of the endangered plant *Primula kwangtungensis* (Primulaceae). *Sci. Data*, **13**, 212.
- 24 39. Lex, A., Gehlenborg, N., Strobel, H., Vuilleumot, R., and Pfister, H. 2014, UpSet:
25 Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.*, **20**, 1983–92.
- 26 40. Pflug, J. M., Holmes, V. R., Burrus, C., Johnston, J. S., and Maddison, D. R. 2020,
27 Measuring Genome Sizes Using Read-Depth, k-mers, and Flow Cytometry:
28 Methodological Comparisons in Beetles (Coleoptera). *G3*, **10**, 3047–60.
- 29 41. Takamura, T. 2007, *Cyclamen Flower Breeding and Genetics*. Springer Netherlands,
30 Dordrecht, pp. 459–78.
- 31 42. Michael, T. P., and VanBuren, R. 2015, Progress, challenges and the future of crop
32 genomes. *Curr. Opin. Plant Biol.*, **24**, 71–81.
- 33 43. Xia, E., Tong, W., Hou, Y., et al. 2020, The reference genome of tea plant and resequencing
34 of 81 diverse accessions provide insights into its genome evolution and adaptation. *Mol.*
35 *Plant*, **13**, 1013–26.
- 36 44. Chen, D., Yang, T., Chen, H., et al. 2025, Largest genome assembly in Brassicaceae:

- 1 retrotransposon-driven genome expansion and karyotype evolution in *Matthiola incana*.
- 2 *Plant Biotechnol. J.*, **23**, 4109–25.
- 3 45. Aida, R., Hirose, Y., Kishimoto, S., and Shibata, M. 1999, *Agrobacterium*
- 4 *tumefaciens*–mediated transformation of *Cyclamen persicum* Mill. *Plant Sci.*, **148**, 1–7.
- 5 46. Terakawa, T., Yamamura, T., and Murayama, T. 2008, Improvement of regeneration and
- 6 transformation systems for *Cyclamen persicum* using somatic embryo culture. *Plant*
- 7 *Biotechnol. (Tsukuba)*, **25**, 77–80.
- 8 47. Tanaka, Y., Yamamura, T., Oshima, Y., et al. 2011, Creating ruffled flower petals in
- 9 *Cyclamen persicum* by expression of the chimeric cyclamen TCP repressor. *Plant*
- 10 *Biotechnol. (Tsukuba)*, **28**, 141–7.
- 11 48. Charlesworth, D., and Willis, J. H. 2009, The genetics of inbreeding depression. *Nat. Rev.*
- 12 *Genet.*, **10**, 783–96.

13

1 **Figure legends**

2 **Figure 1.** Estimated genome size of the *Cyclamen persicum*, based on *k*-mer analysis (*k* = 17),
3 with the given multiplicity values.

4 **Figure 2.** Contig length distribution for the top 95% of the genome assembly.

5 Black, gray, and white bars represent contigs with telomeric motifs at both ends, at one end, or
6 with no detected motifs, respectively.

7 **Figure 3.** Comparative orthogroup analysis among *Cyclamen persicum* and five representative
8 eudicot species. Shared and species-specific orthogroups were identified using OrthoFinder and
9 visualized as an UpSet plot. Numbers indicate orthogroups shared among specific combinations
10 of species.

11

1 **Table 1.** Assembly statistics of the *Cyclamen persicum* genome

	Contig sequences (CPE_r1.0)
Total size (bp)	1,482,524,085
No. of sequences	126
N50 length (bp)	52,300,986
L50	12
Gap length (bp)	0
No. of genes	40,223

2

1 **Table 2.** Repetitive sequences in the cyclamen genome

Repeat type	No. of repeats	Total length (bp)	Relative proportion (%)
SINEs	5,876	535,801	>0.0
LINEs	32,393	18,568,281	1.3
LTR elements	395,762	624,119,859	42.1
DNA transposons	85,726	40,404,166	2.7
Small RNA	2,776	4,042,171	0.3
Satellites	8,922	4,314,249	0.3
Simple repeats	226,076	13,643,877	0.9
Low complexity	25,922	1,267,119	0.1
Unclassified	1,154,127	520,501,486	35.1

2

1 **Table 3.** Completeness of the genome assembly and predicted genes

BUSCOs	Genome assembly (%)	Predicted genes (%)
Complete	98.1	95.7
Single-copy complete	90.2	88.2
Duplicated complete	7.9	7.5
Fragmented	0.7	2.4
Missing	1.2	1.9

2





