scientific data

Check for updates

OPEN

DATA DESCRIPTOR

# Chromosome-level genome assembly of Korean holoparasitic plants, *Orobanche coerulescens*

Bongsang Kim[1,2] ✉, So Yun Jhang[2,3], Bomin Koh[1,2], Soonok Kim[4], Won-Jae Chi[4], Jeong-Mi Park[4], Chae Eun Lim[4], Yoonjee Hong[4], Heebal Kim[1,2,3], Jaewoong Yu[2] ✉ & Seoae Cho[2]

***Orobanche coerulescens*** **is a parasitic plant that cannot complete its life cycle without a host and is incapable of photosynthesis. The habitats of *O. coerulescens* span the coasts of Korea and its volcanic islands, Ulleungdo and Dokdo. Those on the volcanic islands exhibit morphological differences and have distinct hosts compared to those on the peninsula. The family of Orobanchaceae, encompassing both autotrophic and parasitic species, serves as a model for evolutionary studies of parasitic states. However, there are limited genome assemblies for the *Orobanche* genus. In our study, we produced approximately 100x ONT long reads to construct a chromosome-level genome of *O. coerulescens*. The resulting assembly has a total size of 3,648 Mb with an N50 value of 195 Mb, and 82.0% of BUSCO genes were identified as complete. Results of the repeat annotation revealed that 86.3% of the genome consisted of repeat elements, and 29,395 protein-coding genes were annotated. This chromosome-level genome will be an important biological resource for conserving biodiversity and further understanding parasitic plants.**

## Background & Summary

*Orobanche coerulescens*, a member of the Orobanchaceae family, is a holoparasitic plant with an obligate life cycle. Approximately 4,750 species of angiosperms are parasitic, relying wholly or partially on their host plants for water and nutrients to sustain their lives[1]. Parasitic plants can be classified based on their reliance on hosts[2–4]. Obligate parasites necessitate a host to complete their life cycle, while facultative ones do not. Hemiparasites are capable of photosynthesis but obtain water, minerals, and nutrients from their hosts. On the other hand, holoparasites lack chlorophyll and derive their fixed carbon from their hosts, presenting non-green colours. The Orobanchaceae family encompasses diverse autotrophic genera as well as hemiparasitic and holoparasitic lineages, providing an ideal model for evolutionary studies given its comprehensive range of parasitic states[5–7].

While *O. coerulescens* is widely distributed across Eurasia, it is classified as scarce and endangered in Central Europe[8]. Likewise, there is increasing concern about its risk of extinction in Korea[9]. Notably, the species is found on the volcanic islands of Ulleungdo and Dokdo in Korea, exhibiting a distinct geographical niche compared to its presence on the Korean peninsula[10] (Fig. 1). Interestingly, the plants on these islands are glabrous and tend to parasitize on *Artemisia japonica* rather than *Artemisia capillaris*, distinguishing them from those on the mainland counterparts[9]. Despite this ecological specialization, *O. coerulescens* faces a precarious existence due to its limited habitat range and susceptibility to external disturbances, leading to potential declines[9]. The unique geographic and ecological circumstances of *O. coerulescens* found on Ulleungdo and Dokdo Islands underscore the importance of our comprehensive genomic analysis, which aims to understand the evolution of parasitic plants and conserve this unique biodiversity.

To understand the phylogeny of Orobanchaceae and genome evolution mechanisms in parasitic plants, various studies have been conducted. These focus on aspects such as reductive evolution under relaxed selection and structural rearrangement of the genome. Specifically, several studies have targeted on constructing

[1]Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea. [2]eGnome, Inc, Seoul, Republic of Korea. [3]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea. [4]National Institute of Biological Resources, Incheon, Republic of Korea. ✉e-mail: kimbongsang1@gmail.com; jwyu@ungene.co.kr

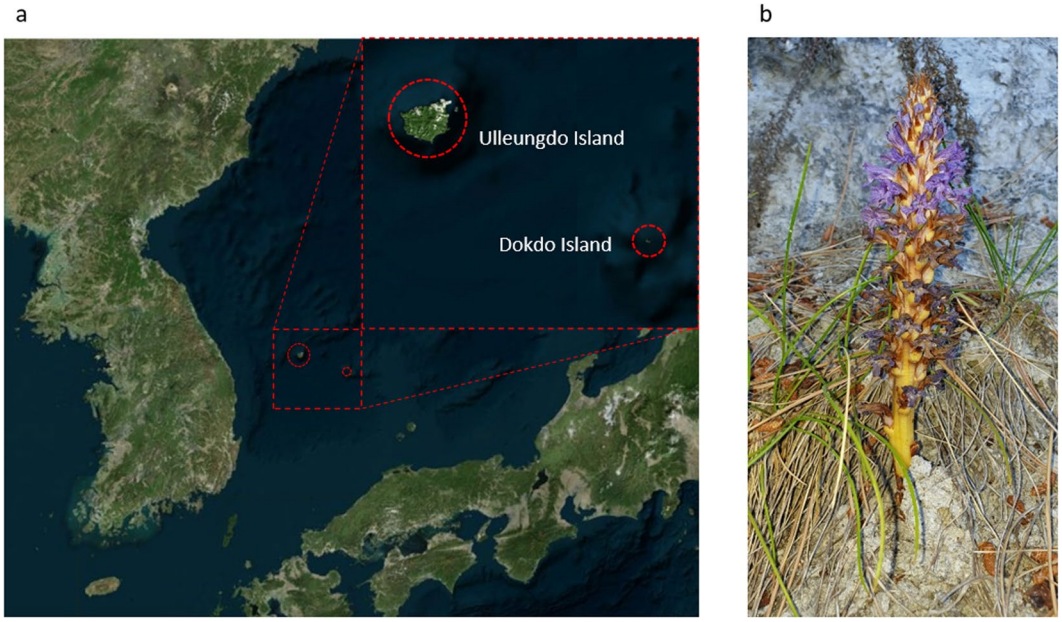


**Fig. 1** Habitats and morphology of *Orobanche coerulescens* on Korean volcanic islands. (**a**) Habitats of *O. coerulescens* on Korean volcanic islands. (**b**) Glabrous type of *O. coerulescens* collected from Ulleungdo Island.

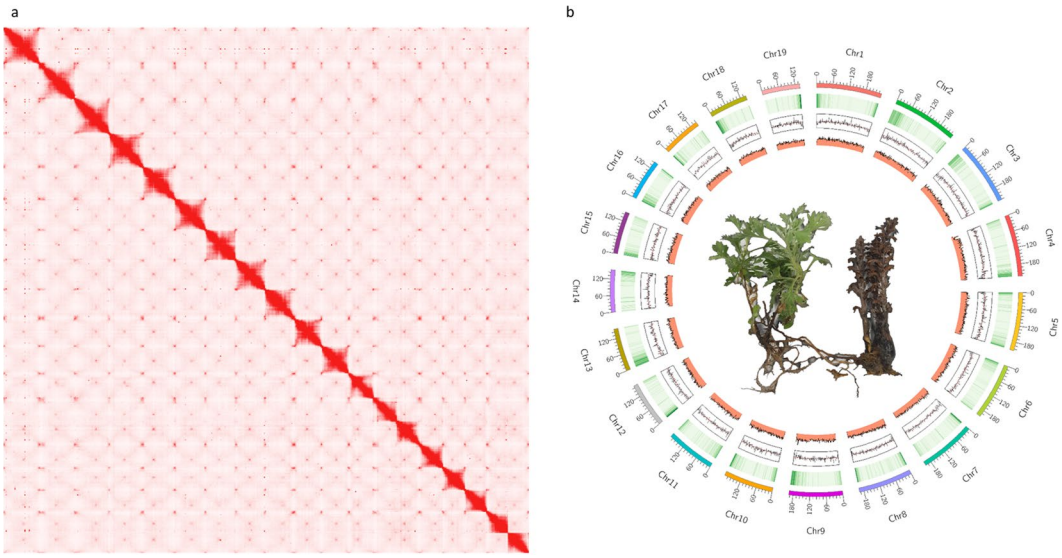


**Fig. 2** Contact map and overview of the *O. coerulescens* genome. (**a**) Contact map of 19 chromosome-level genome. (**b**) Circos plot features from the outermost to the innermost circles indicate gene density, GC content, and density of repeat elements. The figure depicts the host plant, *Artemisia japonica*, and the holoparasitic plant, *O. coerulescens*.

the chloroplast genome of Orobanchaceae[11–13]. However, given the limited scope of these studies to chloroplast genomes, there is a compelling need for complementary research focused on the nuclear genome.

With the ongoing advancements in sequencing technologies, we utilized Oxford Nanopore Technologies (ONT), Illumina, and Hi-C methods to assemble a chromosomal-level genome of *O. coerulescens* (Fig. 2). The final assembly consists of 19 chromosome-level sequences and spans a total length of 3,648 Mb, with an N50 value of 195 Mb. Repeat sequences were found to constitute 86.3% of the entire genome. A total of 29,395 protein-coding genes were annotated with an average length of 4,849 bp, composed of 159,445 exons. This high-quality genome will serve as a valuable genetic resource for understanding the parasitic plant evolutionary mechanisms within Orobanchaceae and for conserving the unique biodiversity of endemics on Korean volcanic islands.

## Methods

**Sampling and sequencing.** *O. coerulescens* was collected from Ulleungdo Island, Korea (37.544826 N, 130.908320E) (Fig. 1). All experimental procedures were conducted in compliance with national and international guidelines. The plant materials used was sourced from adult plants and adhered to with Korean guidelines.

| | Assembly statistics |
|---|---|
| Genome size (bp) | 3,648,123,009 |
| Number of chromsome-scale sequences | 19 |
| Number of sequences | 1,943 |
| Minimum sequence length (bp) | 1,000 |
| N50 (bp) | 195,125,857 |
| Maximum sequence length (bp) | 237,001,669 |
| GC content of the genome (%) | 41.04% |
| | **BUSCO analysis** |
| Complete | 1,323 (82.0%) |
| Complete and single copy | 1,267 (78.5%) |
| Complete and duplicated | 56 (3.5%) |
| Fragmented | 21 (1.3%) |
| Missing | 270 (16.7%) |

**Table 1.** Genome assembly statistics.

The stem of *O. coerulescens* was used to obtain high molecular weight (HMW) DNA following a nuclei isolation method[14] and genomic DNA was extracted using a modified cetyltrimethylammonium bromide (CTAB) protocol[15] in the presence of 2% polyvinylpyrrolidone (1% of molecular weight 10,000 and 1% of molecular weight 40,000) (Sigma-Aldrich, Burlington, MA, USA). For Nanopore sequencing, short genomic fragments were removed using a Short Read Eliminator Kit (Circulomics, Baltimore, MD, USA) and the library was prepared using the ONT 1D ligation Sequencing kit (SQK-LSK109, Oxford Nanopore Technologies, Oxford, UK) with the native barcoding expansion kit (EXPNBD104) in accordance with the manufacturer's protocol. The library was loaded onto a MinION flow cell FLO-MIN106 (Oxford Nanopore Technologies) and PromethION flow cell FLO-PRO002 (Oxford Nanopore Technologies) and sequencing was run on a MinION MK1b and PromethION. A total of long-read DNA data, 354,136,057,138 bp in length, was generated with an N50 value of 33,824 bp and a sequencing depth of 97x. To generate short high-quality sequences, we also used an Illumina platform with TruSeq DNA PCR-Free DNA library (Illumina, San Diego, CA, USA) and then paired-end DNA library was sequenced in the Illumina NovaSeq. 6000 with the length of 151 bp and the depth of 82x (Table S1).

We additionally employed Hi-C data for chromosome-level genome assembly. The Hi-C library was constructed as follows. Ground flower, stem, and root were mixed with 1% formaldehyde for fixing chromatin, and then the nuclei were isolated following a nuclei isolation method[14]. Fixed chromatin was digested with HindIII-HF (New England BioLabs), and we filled the 5′ overhangs with nucleotides and biotin-14-dCTP (Invitrogen) and ligated free blunt ends. After ligation, we purified DNA and removed biotin from unligated DNA ends. Fragmentation and size selection were performed to shear the Hi-C DNA. Hi-C library preparation was performed using the ThruPLEX® DNA-seq Kit (Takara Bio USA, Inc, Mountain View, CA, USA) and sequenced in Illumina NovaSeq. 6000 (Illumina) with the length of 151 bp paired-end reads. The Hi-C data were sequenced in 212,809,764,588 bp with 1,409,336,338 reads (Table S1).

Total RNA was isolated using TRIzol Reagent (Invitrogen) from three tissues of the same *O. coerulescens*-flower, stem, and root-following the manufacturer's protocol. Messenger RNA (mRNA) was isolated using the Magnosphere™ UltraPure mRNA purification kit (Takara) according to the manufacturer's instructions. The cDNA library was prepared using the cDNA-PCR Sequencing Kit (SQK-PCS109, Oxford Nanopore Technologies) with the PCR Barcoding Kit (SQK-PBK004, Oxford Nanopore Technologies) in accordance with the manufacturer's protocol. The library was loaded onto the MinION flow cell FLO-MIN106 (Oxford Nanopore Technologies) and then sequenced on MinION Mk1b. We also used an Illumina platform data using TruSeq Stranded mRNA Prep kit for constructing cDNA library and Illumina NovaSeq. 6000 (Illumina) with the length of 101 bp paired-end reads. The results of RNA sequencing data were described in Table S1.

**Chromosome level genome assembly.** *De novo* assembly was performed using a combination of NextDenovo v2.4.0[16] and NextPolish[17] with default parameters, utilizing ONT long reads and Illumina short reads. To obtain a high-quality chromosome-level genome, we applied the Juicer 1.6[18] and 3D-DNA v180419 pipeline[19] in diploid mode with default parameters. The assembly was further refined and manually curated using a contactmap generated with Juicebox v1.13.01[20]. The final assembly comprised 19 chromosomes, spanning a total length of 3,648,003,138 bp and featuring an N50 value of 195,125,857 bp (Fig. 2 and Table 1).

**Repeat and protein gene annotation.** Repeat annotation in the *O. coerulescens* genome was conducted through two different algorithms: *de novo* and homology-based methods. For *de novo* repeat identification, a custom repeat element library was constructed using RepeatModeler 2.0.3[21] with the long terminal repeat (LTR) structural discovery pipeline option, as well as RECON 1.08[22], RepeatScout 1.0.6[23] and TRF 4.09[24], utilizing the RMBLAST 2.10.0 search engine. Additionally, MITETracker[25] was used to detect miniature inverted-repeat transposable element (MITE) repeat sequences. This custom library was then augmented with data from Repbase 20181026[26] and Dfam 3.6[27] databases to identify repeat sequences homologous to previously known repeat elements. Based on this comprehensive library, RepeatMasker 4.1.4[28] was applied to annotate the repeat elements. In the constructed genome, a total of 3,147,830,327 bp, accounting for 86.29%, was composed of repeat sequences.

| Type | # of elements | length | % of sequence |
|---|---|---|---|
| Retroelements | 2,503,976 | 2,087,694,629 | 57.23% |
| SINEs | 7,260 | 872,464 | 0.02% |
| LINEs | 81,189 | 49,262,217 | 1.35% |
| LTR elements | 2,415,527 | 2,037,559,948 | 55.85% |
| DNA transposons | 459,207 | 201,508,109 | 5.52% |
| Rolling-circles | 18,244 | 10,291,521 | 0.28% |
| Small RNA | 8,102 | 5,234,904 | 0.14% |
| Satellites | 187 | 17,447 | 0.00% |
| Simple repeats | 291,412 | 13,639,164 | 0.37% |
| Low complexity | 38,213 | 2,129,551 | 0.06% |
| Unclassified | 1,994,585 | 827,315,002 | 22.68% |
| Total | 5,313,926 | 3,147,830,327 | 86.29% |

**Table 2.** Statistics of repetitive elements.

| Category | Value |
|---|---|
| Number of total genes | 29,395 |
| Average gene Length | 4,849 bp |
| Number of total exons | 159,445 |
| Number of mono-exon genes | 4,203 |
| Number of multi-exon genes | 25,192 |
| Average exon length | 204 bp |
| Number of total introns | 145,775 |

**Table 3.** Statistics of protein coding gene annotation.

Of these, retroelements such as short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), and LTRs accounted for 0.02%, 1.35%, and 55.85%, respectively, while DNA transposons constituted 5.52% (Table 2).

To annotate protein-coding genes, we first assembled transcriptomes by hybrid approach of long reads and short reads from three tissues. Reads from Illumina and Nanopore were trimmed using Trimmomatic v0.39[29] and Porechop v0.24[30], respectively. The cleaned Illumina reads were then mapped to the *O. coerulescens* genome using STAR and assembled alongside the cleaned Nanopore reads through StringTie v2.2.1[31]. Extracted transcriptomes were processed with TransDecoder v5.5.0[28] using default settings. Gene structure prediction was carried out using the MAKER v3.01.04[32] pipeline, which integrates three different methods: *de novo*-based, transcript evidence-based, and homology-based. *De novo*-based method employed Augustus v3.4.0[33,34], BRAKER v2.1.6[35–39], GeneMark-ET v3.62[40], and SNAP v2.51.7[41] for gene model annotation. Transcript evidence-based method utilizing Exonerate applied to expressed sequence tags (EST) data, while homology-based method sourced data for *Sesamum indicum* (GCF_000512975.1) and *Striga asiatica* (GCA_008636005.1) from National Center for Biotechnology Information (NCBI). Using default settings, MAKER merged the outputs from these three methods to generate weighted consensus gene structures, resulting in the acquisition of 29,395 gene composed of 159,445 exons. The annotated genes included 4,203 mono-exon genes and 25,192 multi-exon genes, with an average gene length of 4,849 bp (Table 3).

## Data Records

The genome assembly was deposited with Genbank accession number GCA_033398695.1[42].

The whole genome sequencing DNA data were deposited with accession number SRR26425285, SRR26465287, SRR26492730, SRR26538103, SRR26538104, SRR26542269, SRR26588572, SRR26588573, SRR26588574, SRR26588624, SRR26588625, SRR26622679, SRR26661892 and SRR26661893 under Sequence Read Archive SRP467180[43].

The transcriptome data of 3 tissues were deposited with accession numbers SRR26426871, SRR26426870, SRR26426869, SRR26503964, SRR26503965 and SRR26503966 under Sequence Read Archive SRP467180[43].

The Hi-C data were deposited with accession number SRR26462611, SRR26522788, SRR26522789 and SRR26622680 under Sequence Read Archive SRP467180[43].

All results from genome annotation are available in figshare[44].

## Technical Validation

**Evaluating genome assembly.** In order to assess the quality of chromosome-level genome of *O. coerulescens*, Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.2.1[45] analysis was performed using the embryophyta_odb10 database, which includes 1,614 single-copy genes. The BUSCO analysis revealed 1,267 complete single-copies and 56 duplicates, accounting for 82.0% completeness, along with 21 fragmented and 270 missing genes (Table 1). Additionally, we carried out BUSCO analysis on the protein-coding genes. Illumina short reads

were subsequently realigned to the genome assembly using BWA v0.7.17[46], yielding a mapping rate of 99.32%. We measured the quality vale (QV) and completeness using Merqury v1.3[47] with k-mer database, resulting in QV value of 29.03 and 92.0% respectively.

**Gene prediction and annotation validation.** To evaluate the gene annotation, protein sequences were searched against reference sequences from *S. lycopersicum*, *A. thaliana*, as well as the NCBI plant RefSeq and the Swiss-Prot databases. The BLASTx v2.8.1[48] was utilized with an e-value threshold below 1e-05 and a query coverage threshold above 40%. As a result, out of a total of 29,395 sequences, successfully matches were found for 24,706 sequences against *S. lycopersicum*, 23,590 sequences against *A. thaliana*, and 29,699 sequences against the combined NCBI and Swiss-Prot databases. Further annotation for motifs, domains and Gene Ontology (GO) was performed using InterProSCan v5.60–92[49]. This resulted in the identification of domains and motifs in 24,252 proteins according to the InterPro database, while 16,681 proteins were annotated with GO terms.

## Code availability

The data analyses were performed according to the manuals and protocols provided by the developers of the respective bioinformatics tools and no custom code was used in the execution of this study. All software and codes used in this work are publicly accessible, and their corresponding versions are specified in Methods section.

Received: 13 November 2023; Accepted: 2 April 2024;
Published online: 02 July 2024

## References

1. Nickrent, D. L. Parasitic angiosperms: how often and how many? *Taxon* **69**, 5–27 (2020).
2. Heide-Jørgensen, H. in *Parasitic flowering plants* (Brill, 2008).
3. Joel, D. M., Gressel, J., Musselman, L. J. & York, c. s. N. Parasitic Orobanchaceae. *Parasitic mechanisms* (2013).
4. Westwood, J. H., Yoder, J. I., Timko, M. P. & dePamphilis, C. W. The evolution of parasitism in plants. *Trends Plant Sci* **15**, 227–235, https://doi.org/10.1016/j.tplants.2010.01.004 (2010).
5. McNeal, J. R., Bennett, J. R., Wolfe, A. D. & Mathews, S. Phylogeny and origins of holoparasitism in Orobanchaceae. *Am J Bot* **100**, 971–983, https://doi.org/10.3732/ajb.1200448 (2013).
6. Schneider, A. C. & Moore, A. J. Parallel Pleistocene amphitropical disjunctions of a parasitic plant and its host. *Am J Bot* **104**, 1745–1755, https://doi.org/10.3732/ajb.1700181 (2017).
7. Li, X., Feng, T., Randle, C. & Schneeweiss, G. M. Phylogenetic Relationships in Orobanchaceae Inferred From Low-Copy Nuclear Genes: Consolidation of Major Clades and Identification of a Novel Position of the Non-photosynthetic Orobanche Clade Sister to All Other Parasitic Orobanchaceae. *Front Plant Sci* **10**, 902, https://doi.org/10.3389/fpls.2019.00902 (2019).
8. Piwowarczyk, R. Revised distribution and phytosociological data of Orobanche coerulescens Stephan in Willd.(Orobanchaceae): Poland in relation to Central Europe. *Biodiversity: Research Conservation* **26**, 61–72 (2012).
9. Lee, W. *et al*. Morphological variation and aspects of the geographic distribution of Orobanche coerulescens Stephan ex Willd. (Orobanchaceae) on Ulleung-do and Dok-do Islands. *Korean Journal of Plant Taxonomy* **46**, 405–412 (2016).
10. Kim, B. *et al*. Distinctive origin and evolution of endemic thistle of Korean volcanic island: Structural organization and phylogenetic relationships with complete chloroplast genome. *PLoS One* **18**, e0277471, https://doi.org/10.1371/journal.pone.0277471 (2023).
11. Park, J. M., Manen, J. F., Colwell, A. E. & Schneeweiss, G. M. A plastid gene phylogeny of the non-photosynthetic parasitic Orobanche (Orobanchaceae) and related genera. *J Plant Res* **121**, 365–376, https://doi.org/10.1007/s10265-008-0169-5 (2008).
12. Qin, L. *et al*. The complete chloroplast genome of Striga asiatica (L.) Kuntze 1891 (Orobanchaceae), a hemiparasitic weed from Guangxi China. *Mitochondrial DNA B Resour* **8**, 497–500, https://doi.org/10.1080/23802359.2023.2197089 (2023).
13. Zhou, T. *et al*. The Complete Chloroplast Genome of Euphrasia regelii, Pseudogenization of ndh Genes and the Phylogenetic Relationships Within Orobanchaceae. *Front Genet* **10**, 444, https://doi.org/10.3389/fgene.2019.00444 (2019).
14. Zhang, M. *et al*. Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. *Nat Protoc* **7**, 467–478, https://doi.org/10.1038/nprot.2011.455 (2012).
15. Porebski, S., Bailey, L. G. & Baum, B. R. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant molecular biology reporter* **15**, 8–15 (1997).
16. Hu, J. *et al*. An efficient error correction and accurate assembly tool for noisy long reads. *bioRxiv*, 2023.2003. 2009.531669 (2023).
17. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255, https://doi.org/10.1093/bioinformatics/btz891 (2020).
18. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
19. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).
20. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
21. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
22. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12**, 1269–1276, https://doi.org/10.1101/gr.88502 (2002).
23. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**(Suppl 1), i351–358, https://doi.org/10.1093/bioinformatics/bti1018 (2005).
24. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580, https://doi.org/10.1093/nar/27.2.573 (1999).
25. Crescente, J. M., Zavallo, D., Helguera, M. & Vanzetti, L. S. MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinformatics* **19**, 348, https://doi.org/10.1186/s12859-018-2376-y (2018).
26. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11, https://doi.org/10.1186/s13100-015-0041-9 (2015).
27. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* **12**, 2, https://doi.org/10.1186/s13100-020-00230-y (2021).
28. Smit, A., Hubley, R. & Green, P. (2015).
29. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120, https://doi.org/10.1093/bioinformatics/btu170 (2014).
30. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom* **3**, e000132, https://doi.org/10.1099/mgen.0.000132 (2017).

31. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
32. Cantarel, B. L. *et al*. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**, 188–196, https://doi.org/10.1101/gr.6743907 (2008).
33. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644, https://doi.org/10.1093/bioinformatics/btn013 (2008).
34. Stanke, M., Schoffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62, https://doi.org/10.1186/1471-2105-7-62 (2006).
35. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stromberg, M. P. & Marth, G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692, https://doi.org/10.1093/bioinformatics/btr174 (2011).
36. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60, https://doi.org/10.1038/nmeth.3176 (2015).
37. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769, https://doi.org/10.1093/bioinformatics/btv661 (2016).
38. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-Genome Annotation with BRAKER. *Methods Mol Biol* **1962**, 65–95, https://doi.org/10.1007/978-1-4939-9173-0_5 (2019).
39. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
40. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* **42**, e119, https://doi.org/10.1093/nar/gku557 (2014).
41. Leskovec, J. & Sosic, R. SNAP: A General Purpose Network Analysis and Graph Mining Library. *ACM Trans Intell Syst Technol* **8**, https://doi.org/10.1145/2898361 (2016).
42. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_033398695.1 (2023).
43. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP467180 (2023).
44. Kim, B. *et al*. Annotation result of Orobanche coerulescens. *figshare* https://doi.org/10.6084/m9.figshare.c.6888202.v1 (2023).
45. Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).
46. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv* (2013).
47. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
48. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, https://doi.org/10.1186/1471-2105-10-421 (2009).
49. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).


## Acknowledgements

This work was supported by a grant from the National Institute of Biological Resources (NIBR), funded by the Ministry of Environment (MOE) of the Republic of Korea (NIBR201930201, NIBR202331203). The funders participated in study design and revised the manuscript.

## Author contributions

Soonok Kim, Jaewoong Yu, and Seoae Cho designed the project; Won-Jae Chi, Chae Eun Lim and Yoonjee Hong conceived ideas and prepared funding; Bomin Koh collected the sample and conducted laboratory experiments; Bongsang Kim constructed the assembly and performed annotation; Bongsang Kim and So Yun Jhang wrote the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03207-1.

**Correspondence** and requests for materials should be addressed to B.K. or J.Y.