Chromosome-level reference genome assemblies of five Brassicaceae species inhabiting challenging ultramafic substrates

Mahnaz Nezamivand-Chegini

nezamivm@natur.cuni.cz

Charles University

Miloš Duchoslav

Charles University

Raúl Wijfjes

Heinrich Heine University Düsseldorf

Gabriela Šrámková

Charles University

Jana Nosková

Charles University

Vít Bureš

Charles University

Tereza Koberová

Charles University

Terezie Mandáková

Masaryk University

Tomica Mišljenović

University of Belgrade

Ksenija Jakovljević

University of Belgrade

Panayiotis G. Dimitrakopoulos

University of the Aegean

Stanislav Španiel

Slovak Academy of Sciences

Bruno Huettel

Max Planck Institute for Plant Breeding Research

Martin A. Lysak

Masaryk University

Levi Yant

Swedish University of Agricultural Sciences

Korbinian Schneeberger

Heinrich Heine University Düsseldorf

Filip Kolář

Charles University

data-descriptor

Keywords:

Posted Date: July 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8960457/v1

Additional Declarations: No competing interests reported.

Abstract

We present chromosome-scale genome assemblies for five Brassicaceae species: *Aethionema saxatile*, *Cardamine glauca*, *Erysimum linariifolium*, *Noccaea praecox*, and *Odontarrhena muralis*. These species inhabit challenging ultramafic (serpentine) substrates, and all come from the Balkan peninsula, a biodiversity hotspot of European flora that stood aside from large genomic initiatives and investigations. The species represent two subfamilies - Aethionemoideae (tribe Aethionemeae) and Brassicoideae, and three Brassicoideae supertribes: Camelinodae (*C. glauca*, *E. linariifolium)*, Brassicodae (*N. praecox*) and Arabodae (*O. muralis*). Utilising PacBio HiFi sequencing and Hi-C scaffolding, we produced nearly complete telomere-to-telomere assemblies with scaffold N50 values between 20 and 34 Mb. Quality assessments indicate high completeness, with BUSCO scores exceeding 95%. Annotation of repetitive elements revealed that the species vary in their repeat content. These assemblies, accompanied by protein-coding gene annotations with orthology and functional information, expand genomic resources for comparative and evolutionary studies within the family and provide foundational resources for investigating adaptation towards challenging environments in an understudied European diversity hotspot.

Background & Summary

The Brassicaceae family, comprising over 4,000 species, is one of the most extensively studied plant groups due to its remarkable ecological and morphological diversity, generally small and accessible genomes and presence of several crop and model species, including *Arabidopsis thaliana*[1]. Phylogenetically, the family is divided into two subfamilies: the early-diverging Aethionemoideae, which includes the genus *Aethionema* (tribe Aethionemeae), and the much larger Brassicoideae, which accounts for approximately 98.6% of species diversity. Within Brassicoideae, recent phylogenomic studies have established five supertribes that correspond to distinct evolutionary lineages: Camelinodae (Lineage I), Brassicodae (Lineage II), Hesperodae (Lineage III), Arabodae (Lineage IV), and Heliophilodae (Lineage V)[2,3].

The family provided valuable model systems for understanding genetic underpinnings of adaptation to a broad array of challenging environments, including alpine, saline, metal-contaminated or arctic habitats[4–7]. The five species sequenced in this study represent adaptation to another challenge: naturally toxic ultramafic soils. Ultramafic (sometimes narrowly termed serpentine) soils, characterised by skewed elemental ratios (particularly Ca/Mg), high metal content and shortage of essential macronutrients[8,9], represent a suitable model for studying plant adaptation because the selective factors are exceptionally strong and replicated across spatially restricted outcrops[10]. In addition, two species, *Noccaea praecox* and *Odontarrhena muralis*, are known to hyperaccumulate Ni[11], exhibiting a potential for applications in phytomining and phytoremediation[12]. The five species cover a broad phylogenetic range within the Brassicaceae, including four major clades within the family, and all come from the Balkans - an understudied hotspot of European diversity[13]. In this context, chromosome-scale genome assemblies

provide valuable tools for understanding evolutionary history and dissecting the genetic architecture of adaptive traits across diverse Brassicaceae species[14]. These assemblies are not only essential for addressing evolutionary questions related to comparative genomics and adaptation studies, but also for practical applications such as crop improvement and conservation biology.

Methods

Sampling.

The seeds of all five species were sampled in Serbia (*E. linariifolium, N. praecox, O. muralis*) and Greece (*Ae. saxatile, C. glauca*, see Supplementary Table 1 for locality details). Fresh tissue was collected from plants cultivated in an experimental greenhouse at the Department of Botany, Charles University, snap-frozen in liquid nitrogen, and stored at − 80°C until analysis.

Chromosome counting and flow cytometry.

Chromosome numbers were cytologically validated using root tip meristematic tissue from cultivated plants originating from the same population as the sequenced individuals, following[15]. Root tips were incubated in ice-cold distilled water for 24 h, fixed overnight at 4°C in Carnoy's fixative (acetic acid: ethanol, 1:3), and stored in 70% ethanol at 4°C. Mitotic chromosome spreads were prepared by squash technique after partial digestion with a pectolytic enzyme mixture[15]. Chromosomes were stained with 4',6-diamidin-2-fenylindole (DAPI) and imaged using a Zeiss Axioimager epifluorescence microscope equipped with a CoolCube camera (MetaSystems). Representative mitotic spreads are shown in Fig. 1, and haploid chromosome numbers are listed in Table 1.

Genome sizes were determined by flow cytometry following the two-step procedure with Otto buffers[16,17]. Approximately 0.5 cm^2 of intact leaf tissue and a corresponding amount of internal reference standard were co-chopped in 0.5 mL of Otto I buffer. The suspension was filtered through a 42µm nylon mesh and incubated for 15 min at room temperature. Subsequently, 1 mL of Otto II buffer supplemented with propidium iodide and RNase IIA (both at 50 µg/mL) was added. Two internal standards were used according to the expected genome size of the samples: *Carex acutiformis* (2C = 0.82 pg)[18] or *Solanum pseudocapsicum* (2C = 2.59 pg)[19]. Samples were analyzed using a Partec CyFlow SL flow cytometer (Sysmex Partec GmbH, Görlitz, Germany) equipped with a 532 nm green solid-state laser. The fluorescence intensity of 3,000 nuclei was recorded for each run, and each plant was measured three times on different days to ensure the reliability of the results. Genome sizese were calculated as the ratio of the mean positions of the G1 peaks of the sample and the internal standard, multiplied by the known genome size of the standard. Haploid genome sizes are shown in Table 1.

Library construction & sequencing.

For *E. linariifolium*, and *N. praecox*, high-molecular-weight (HMW) DNA from fresh-frozen material using a NucleoBond HMW DNA kit (Macherey-Nagel) following manufacturer's protocol. HMW DNA quality was

confirmed using Femto pulse system (Agilent) and sheared to the appropriate size range (15–20 kb). PacBio HiFi sequencing libraries were constructed using SMRTbell prep kit 3.0 (PacBio). For *C. glauca*, *O. muralis*, and *Ae. saxatile*, HMW DNA was extracted from living plants using a protocol described by Russo *et al.*[20], whereas HMW DNA quality was confirmed by gel electrophoresis and sheared to the appropriate size range (15–20 kb). Novogene (UK) Company Limited constructed PacBio HiFi sequencing libraries. The sequencing for each sample was performed on one full SMRT cell of the PacBio Sequel Revio system (Supplementary Table 2). Due to the small amount of available fresh-frozen or fresh material, Hi-C libraries were prepared from leaves of half-sibling plants using the Arima Genomics® High Coverage Hi-C kit (Carlsbad, California, USA) according to the manufacturer's protocol. Novogene (UK) Company Limited sequenced the Hi-C library on an Illumina NovaSeq 6000 S4 platform (Supplementary Table 3).

To support genome annotation, mRNA from 4–6 tissue types per species were sequenced, including flowers, flower buds, young and mature leaves, stems, and roots. Total RNA was isolated using the NucleoSpin RNA Plus Kit (Macherey-Nagel) according to the manufacturer's protocol, with two modifications: increased volumes of Lysis Buffer LBP (500 µl) and Binding Solution BS (170 µL) were applied. mRNA libraries were generated using poly(A) enrichment (Novogene samples) or using TruSeq Stranded mRNA Library Prep Kit (Macrogen). Sequencing was carried out on an Illumina sequencer with 150-bp paired-end reads, yielding 16.6–91.7 million read pairs per sample (Supplementary Table 4).

Genome size estimation.

HiFi reads were processed using KMC (v3.1.1)[21] for k-mer counting with parameters "-k31 -m64 -cs 2000000", followed by GenomeScope2 (v2.0)[22] to estimate genome characteristics (Table 1, Fig. 2). The haploid genome length estimates ranged from 168.99 to 255.21 Mbp, with repeat content varying between 45.26 and 85.72 Mbp (26.8%-38.9%). These values are consistent with flow cytometric estimates using propidium iodide staining (Table 1). Heterozygosity levels were lowest in *Ae. saxatile* (0.08–0.10%) and highest in *C. glauca* (1.37–1.39%). The minor deviation from 100% of the total k-mer spectrum modelled by GenomeScope2 is attributed to sequencing error k-mers and model fitting uncertainty, as is typical for GenomeScope2 analyses[22].

Table 1
Summary of genome size estimates and cytogenetic characteristics for the five Brassicaceae species (*Aethionema saxatile*, *Cardamine glauca*, *Erysimum linariifolium*, *Noccaea praecox* and *Odontarrhena muralis*).

Properties	***Ae. saxatile***	***C. glauca***	***E. linariifolium***	***N. praecox***	***O. muralis***
Homozygous (aa) (%)	99.91	98.63	98.83	98.67	99.11
Heterozygous (ab) (%)	0.10	1.39	1.17	1.34	0.91
Haploid genome size (Mbp)	255.21	224.48	168.99	209.56	189.99
Genome repeat seq. length (Mbp)	85.72	78.98	45.26	81.51	54.06
Genome unique seq. length (Mbp)	169.48	145.50	123.73	128.05	135.94
Model fit (%)	92.25	91.93	94.10	93.99	90.75
Read error rate (%)	0.46	0.31	0.22	0.24	0.15
Haploid genome size - flow cytometry (Mbp)	280	265	185	245	235
Haploid chromosome number (n)	12	8	7	7	8

Values from GenomeScope2 (k-mer–based) include the estimated proportions of homozygous (aa) and heterozygous (ab) k-mers, haploid genome size, repeat length, unique sequence length, model fit, and inferred sequencing error rate (“read error rate”). “Model fit” indicates how well the GenomeScope2 mixture model explains the observed k-mer spectrum.

Nuclear genome assembly

Hifiasm v0.20[23] was used to generate the primary assemblies (Fig. 3). For all species except *Ae. saxatile*, we ran Hifiasm in HiC-integrated mode with purging to handle heterozygous sequences. However, for *Ae. saxatile*, low heterozygosity suggested that phasing was unnecessary. Therefore, Hifiasm was run in solo mode with no purging (-l0) to avoid mispurging and incorrect phasing. After generating the primary assemblies with Hifiasm, variation in contiguity across species was observed. *C. glauca* had the fewest number of contigs (821), whereas *E. linariifolium* showed the most fragmentation, with around 1910 contigs.

First, to address potential contamination, both Blobtools v4.3.0[24] and FCS-GX (v0.5.5)[25] were used to identify and filter out contaminant sequences. This revealed contamination primarily from *Arthropoda* and some fungal species. Second, to identify plastid and mitochondrial contigs, sequence similarity searches were performed using MMseqs2[26 26] against the NCBI RefSeq organelle genome Database, including plastid and mitochondrial reference sequences[27]. Then, Purge_haplotigs v1.1.3[28] was used to

identify redundant haplotype sequences representing alternative assemblies of the same genomic region from different homologs. Thresholds were set for minimum, median, and maximum coverage to classify contigs accordingly, ensuring that true haploid sequences were retained while duplicates and repetitive regions were flagged for removal. Finally, the contigs flagged as contaminants, haplotigs, junks, and hits to organelle DNA based on the results from the above analyses were filtered out. This filtration significantly reduced the total number of contigs across all species, leading to a substantially reduced number of contigs.

The scaffolding of the retained contigs proceeded by aligning Hi-C paired-end reads to the assembly using Omni-C pipeline[29], and alignments were filtered by mapping quality using SAMtools v1.14[30] (-q 10). Filtered alignments were converted to the BED format using BEDTools v2.26.0[31] for downstream scaffolding and visualisation. Automated scaffolding was performed using SALSA2[32,33], and the resulting assembly was visualised and manually curated using JuiceBox[34] based on HiC contact map. Repeatmasker[35] was run after each manual correction to annotate telomere repeats using the telomeric repeat sequence of *Arabidopsis thaliana* as a repeat library. HiC scaffolding produced chromosome-scale scaffolds for each species. Telomeric repeat sequences were detected at both ends of most scaffolds, whereas scaffold 6 in Ae. Saxatile and scaffolds 2, 3, 5, and 5 in O. muralis lacked telomeric repeats at one or both ends. Additional smaller scaffolds representing unplaced sequences were also kept in the final assemblies.

Finally, the assemblies were polished by correcting sequence errors in regions with reliable read support. HiFi reads were mapped back to the assemblies using Winnowmap2 (v2.03)[36], and genome-wide coverage distributions were generated with mosdepth v0.3.11[37]. For each species, we computed mean coverage in fixed-size windows of 10 kbp and examined the resulting histograms (not shown) to identify the central, well-supported coverage peak. Scaffolds in which all windows fell outside the normal coverage range (i.e., entirely low- or high-coverage) were considered unreliable and were removed from further analysis. The remaining scaffolds, which contained windows within the species-specific normal coverage range, were retained for polishing.

For these retained scaffolds, sequence errors were identified using two complementary variant callers; DeepVariant v1.9.0[38] for small variants (SNPs and short indels < 50 bp) and CuteSV v2.1.1[39] for structural variants (insertions, deletions, and rearrangements > 50 bp). For both callers, only homozygous variants with high allele support were retained. All retained variants were visually inspected in IGV (v2.18.2)[40] to remove false positives. The curated small-variant and SV sets were then combined and applied to correct the assemblies using bcftools consensus, replacing incorrect sequences.

Annotation of repetitive sequences.

De novo repeat annotation was performed for all species using RepeatModeler v2.0.4[41] with LTRStruct enabled. For each genome, the process began with the creation of a custom repeat database using BuildDatabase, followed by the identification and classification of repeat families with RepeatModeler.

The resulting classified repeat sequences were compiled into a species-specific repeat library. For homology-based annotation, RepeatMasker v4.1.2[35] was applied using the custom repeat libraries generated by RepeatModeler. This pipeline allowed for comprehensive identification and classification of repetitive elements, including retroelements, DNA transposons, rolling-circle elements, unclassified repeats, simple repeats and low-complexity regions. Repeat content varied significantly among the five species, with total repeat content ranging from 34.72% in *E. linariifolium* to 62.95% in *Ae. saxatile* (Table 2). Retroelements constituted the largest fraction of repetitive DNA in all genomes, with long terminal repeat (LTR) elements being the most abundant subclass. Among LTR retrotransposons, Ty1/Copia and Gypsy families dominated, although their genomic proportion varied across species. For instance, *C. glauca* contained a higher genomic proportion of Ty1/Copia elements (11.57%) compared to *Ae. saxatile* (2.76%), whereas Gypsy elements were particularly enriched in *Ae. saxatile* (24.02%). In contrast, DNA transposons and LINEs (long interspersed nuclear elements) were consistently less abundant, together contributing less than 5% of the genome in all assemblies. Unclassified repeats represented a substantial component of the repeat landscape, particularly in *Ae. saxatile* (24.84%) and *N. praecox* (23.3%). Simple repeats and low-complexity regions were present at low level across all species, each constituting less than 2% of the genome. Satellite DNA was nearly absent, with only trace amounts detected in *C. glauca, E. linariifolium*, and *N. praecox*.

Table 2
Summary of repeat element content in the genome assemblies of five species (*Aethionema saxatile*, *Cardamine glauca*, *Erysimum linariifolium*, *Noccaea praecox* and *Odontarrhena muralis*).

Category	***Ae. saxatile***	***C. glauca***	***E. linariifolium***	***N. praecox***	***O. muralis***
Total repeat content (%)	62.95	62.10	34.72	52.15	49.33
Retroelements (%)	29.51	40.62	18.96	23.36	28.47
LTR retrotransposons (%)	29.02	39.32	17.52	21.73	26.85
Ty1/Copia (%)	2.76	11.57	11.29	3.47	13.03
Gypsy (%)	24.02	21.83	4.41	15.18	9.06
LINEs (%)	0.50	1.28	1.37	1.63	1.63
DNA transposons (%)	3.31	1.55	0.91	1.97	3.01
Unclassified repeats (%)	24.84	15.58	10.17	23.30	13.85
Satellite DNA (%)	0	0.05	0.02	0.02	0.02
Simple repeats (%)	1.52	1.10	1.17	0.89	1.02
Low-complexity regions (%)	0.30	0.22	0.33	0.22	0.26

Plastid genome assembly and annotation

The plastid genomes (plastomes) of each species were assembled and annotated using OATK v1.0[42] with default parameters for organelle genome assembly. The plastomes were circularised and annotated using OATK's integrated pipeline. To visualise gene organisation, circular maps were generated using OGDRAW v1.3.1[43]. The assembled plastome sizes ranged from 152,691 bp (*N. praecox*) to 154,633 bp (*C. glauca*) (Table 3).

All five plastomes exhibited the canonical quadripartite structure of land plants (Fig. 4). They also shared an identical complement of 114 genes, comprising 80 protein-coding genes, 30 tRNAs, and 4 rRNAs. The 30 tRNA genes represented all standard anticodons for plastid translation. OGDRAW visualisation confirmed proper gene partitioning, with rRNA operons (*rrn16-rrn23-rrn4.5-rrn5*) correctly localised to the IR regions, the *rps12* trans-splicing gene split between the IR and LSC regions, and the *ycf1* and *ycf2* genes flanking the IR/SSC boundaries as expected for canonical angiosperm plastomes[44].

Table 3
Summary of plastid genome assembly statistics and gene content for the five Brassicaceae species (*Aethionema saxatile*, *Cardamine glauca*, *Erysimum linariifolium*, *Noccaea praecox* and *Odontarrhena muralis*).

Ae. saxatile	**Total length (bp)**	**LSC length (bp)**	**SCS length (bp)**	**IRA (bp)**	**IRB (bp)**	**CDS**	**tRNA**	**rRNA**
	154149	**70673**	**17743**	**26463**	**26463**	**85**	**37**	**8**
C. glauca	154633	70666	17834	26414	26414	87	37	6
E. linariifolium	154502	70686	17846	26418	26418	87	37	8
N. praecox	152691	68306	17946	25178	25178	85	38	8
O. muralis	153676	70958	17966	26494	26494	85	36	8

Nuclear genome annotation

As one of the inputs for gene prediction, our RNA-sequencing (RNA-seq) data from the five target species was used (Supplementary Table 5). For *N. praecox*, published RNA-seq data[45] downloaded from ENA (PRJNA984443) was also used. RNA-seq data were trimmed to remove sequencing adapters and low-quality regions using Trim Galore v.0.6.2[46] with paired mode ('--paired') and aligned to the assembled reference genome using HISAT2 v.2.1.0[47] with command line options '-q -t --dta'. The resulting SAM files were converted to BAM files, sorted and merged for each species using SAMtools v.1.14[30].

The protein-coding genes were predicted by BRAKER v.3.0.8[48] with default settings based on our RNA-seq data and on Viridiplantae proteins from OrthoDB v.11 database[49]. The gene models with internal stop codons were removed. To include regions that are transcribed but do not encode proteins (non-coding RNA genes and similar features), transcripts from RNA-seq data was assembled using StringTie

v.2.2.3 with default settings[50] and filtered using BEDTools v.2.30.0[31] to keep only those that had no overlap with protein-coding gene annotations. These elements were included in the annotation as 'ncRNA_gene' and 'ncRNA'. We converted GTF to GFF using AGAT v.1.4.0[51].

To assess reliability of predicted protein-coding genes, they were overlapped with (1) transcripts assembled from RNA-seq used for annotation by StringTie and with (2) protein sequences from three well-annotated Brassicaceae species aligned to the genome using Miniprot v.0.13-r248[52]. Proteomes from *Arabidopsis thaliana* (Araport11, version 20220914 downloaded from https://www.arabidopsis.org), *Arabidopsis lyrata* (GCF_000004255.2 downloaded from GenBank) and *Brassica rapa* (version Brapa_chiifu_v41_gene20230413 downloaded from http://brassicadb.cn) were used. The overlaps were calculated using BEDTools v.2.30.0[31] intersect command with setting '-f 0.9' for intersects with assembled transcripts (the overlap should be at least 90% of predicted gene) and '-f 0.9 -r' (the overlap should be at least 90% of both the predicted gene and the aligned protein). The numbers of predicted protein-coding genes ranged from ~ 25 000 (*Ae. saxatile, C. glauca*) to ~ 30 000 (*E. linariifolium, N. praecox* and *O. muralis*; Table 4). Interestingly, *C. glauca* with the lowest number of genes (24 828) had a similar number of supported genes as other species, so the difference was mainly due to less reliable gene models (Table 4).

Table 4
Number of predicted genes and their proportion supported by assembled transcripts from available RNA-seq data and aligned proteins from three Brassicaceae species (*Arabidopsis thaliana*, *Arabidopsis lyrata* and *Brassica rapa*). For support by aligned proteins, reported are genes supported by protein from at least one of the three species.

Species	**Protein-coding genes**				**Non-coding RNA genes**
	Total	**Supported by assembled transcripts N (%)**	**Supported by aligned proteins N (%)**	**Supported by assembled transcripts or aligned proteins N (%)**	**Total**
Ae. saxatile	25 353	19 189 (76)	18 200 (72)	21 293 (84)	4 458
C. glauca	24 828	20 452 (82)	21 365 (86)	23 307 (94)	8 901
E. linariifolium	30 393	22 625 (74)	22 348 (74)	26 422 (87)	7 271
N. praecox	30 370	21 522 (71)	21 042 (69)	25 080 (83)	13 747
O. muralis	29 940	20 841 (70)	20 261 (68)	24 066 (80)	7 786

Genome architecture and feature distribution analysis

To visualise the genomic architecture and distribution of key genomic features across the five assembled genomes, we generated Circos plots displaying five concentric tracks representing distinct genomic features in 100 kb windows: gene density, repeat content, GC content, centromeric repeats, and telomeric repeat distributions (Fig. 5).

The outermost track represents gene density, showing the number of protein-coding genes per 100 kb window normalised from 0-100. All five species exhibited canonical eukaryotic chromosome architecture with gene-rich euchromatic arms and gene-poor pericentromeric heterochromatin. This pattern is most clearly visible in *C. glauca, E. linariifolium* and *O. muralis*, where gene-sparse pericentromeric regions contrast sharply with gene-dense chromosome arms, while *Ae. saxatile* and *N. praecox* show more gradual transitions between these regions.

The second and third tracks display repeat density and GC content, both as percentages per 100 kb window. Both tracks characteristic inverse correlation with gene density, with repeat-rich, GC-poor pericentromeric heterochromatin contrasting sharply with repeat-poor, GC-rich euchromatic arms. The fourth track represents centromeric repeat density based on TRASH (Tandem Repeat Annotation and Structural Hierarchy) analysis[53]. To comprehensively visualise centromeric organisation, we merged TRASH-identified repeat families for each species; domains 58, 219, 291, 400, and 438 for *Ae. saxatile*, domains 58, 81, 162, 242, 321, 428, 505, and 508 for *C. glauca*, domains 48, 160, 318, and 516 for *E. linariifolium*, domains 199, 478, and 771 for *N. praecox*, and domains 69, 118, 156, 206, 316, 323, and 346 for *O. muralis*. The resulting visualisations revealed heterogeneity in centromeric repeat composition in some scaffolds, suggesting complex tandem repeat organisation characteristic of plant centromeres.

The innermost track displays telomeric repeat density based on TIDK (Telomeric Identification Toolkit, reference)[54] analysis of canonical plant telomeric repeat motif TTTAGGG. Telomeric signals were strongest at chromosome termini, confirming successful assembly to telomeric regions for nearly all scaffolds. Interstitial Telomeric Sequences (ITSs) were also detected throughout the genome. The circos visualisations confirm high-quality genome assemblies with clear differentiation of euchromatic and heterochromatic regions, properly positioned centromeres and telomeric sequences at most scaffold termini, validating the structural integrity of the assemblies.

Orthology between predicted genes and functional annotation

As an additional resource, orthology assignment and functional annotation of protein-coding genes were performed. To assign orthology relationships between genes of different species, OrthoFinder v.2.5.4[55] was run with proteomes of the currently presented species and additional 17 Brassicaceae species with reference genomes and annotations of suitable quality (*Alyssum gmelinii, Arabidopsis arenosa, Arabidopsis lyrata, Arabidopsis thaliana, Arabis alpina, Brassica napus, Brassica oleracea, Brassica rapa, Camelina sativa, Capsella rubella, Cardamine amara, Cardamine hirsuta, Cochlearia excelsa, Conringia*

planisiliqua, Euclidium syriacum, Eutrema salsugineum and *Raphanus sativus*). Additionally, BLAST reciprocal best hits analysis[56] was performed with default values (minimally 70% identity over the aligned region, which should cover at least 50% of the query sequence) and simple BLAST search (blastp 2.16.0+)[57] of each species' proteome against *A. thaliana* proteome. *A. thaliana* orthologues (or other kinds of homologues) to the genes of other species in one of five ways, using the next step only if no *A. thaliana* gene was assigned in the previous one: (1) Orthologues (from "Orthologues" directory) produced by OrthoFinder, (2) BLAST reciprocal best hits, (3) genes from N0 phylogenetic hierarchical orthogroups produced by OrthoFinder, (4) genes from broader orthogroups (from "Orthogroups" directory) produced by OrthoFinder and (5) best hits (based on bitscore) from simple BLAST search (E-value minimally 0.1, identity over the aligned region minimally 70%, aligned region should cover at least 50% of the query sequence). OrthoFinder assigned direct *A. thaliana* orthologues (method 1) to 70–86% of protein coding genes across our focal species, while adding homologues by methods 2–5 increased the number of genes with assigned *A. thaliana* homologue to 85–97% (Table 5). We thus call the *A. thaliana* genes added by the methods 2–5 just "homologues", as their orthology/paralogy status is less clear (most of them are probably out-paralogues).

Functional annotations to the genes were added by running InterProScan v. 5.76–107[58] with protein sequences of each species and by transferring annotations from *A. thaliana* homologues assigned in previous steps. The following sources of functional annotations of *A. thaliana* were used: (1) The Arabidopsis Information Resource (TAIR)[59] Gene aliases, Subcellular predictions, Functional descriptions and GO annotations, version 20241001; (2) AraCyc Pathways v. 20230103 (Hawkins et al., 2025) and (3) UniProt database (data downloaded 2025-11-11)[60]. In case of several *A. thaliana* orthologues/homologues assigned to one gene, functional annotations from all of them were combined. The final set of Gene ontology (GO) terms was combined from three sources - InterProScan (based on sequences from each species), TAIR GO annotations (from *A. thaliana* homologues) and UniProt GO annotations (from *A. thaliana* homologues). In this way, we were able to obtain GO annotations for 91–98% of protein coding genes in our target species, with average number of GO terms per gene between 6.7 and 7.7 (Table 5).

Table 5

Orthology and homology assignment and functional annotation of protein-coding genes. "Orthologues" denote direct orthologues assigned by OrthoFinder, "homologues" denote homologues added by other methods described in Methods. Mean count of gene ontology (GO) terms was calculated from all protein-coding genes of the species, including the ones that have assigned no GO terms.

Species	Protein-coding genes with assigned *A. thaliana* orthologue N (%)	Protein-coding genes with assigned *A. thaliana* orthologue or other homologue N (%)	Protein-coding genes annotated with GO terms N (%)	Mean count of GO terms per protein-coding gene
Ae. saxatile	19 191 (76)	21 742 (86)	21 993 (87)	7.0
C. glauca	21 391 (86)	24 080 (97)	24 258 (98)	7.7
E. linariifolium	23 161 (76)	27 783 (91)	28 219 (93)	7.1
N. praecox	21 419 (71)	26 350 (87)	26 955 (89)	6.8
O. muralis	20 924 (70)	25 569 (85)	26 378 (88)	6.7

Data record

All data generated in this study are publicly available at the National Center for Biotechnology Information (NCBI) under BioProject accession PRJNA1276498. The BioProject contains chromosome-scale genome assemblies and associated raw sequencing datasets for five Brassicaceae species: *Aethionema saxatile*, *Cardamine glauca*, *Erysimum linariifolium*, *Noccaea praecox*, and *Odontarrhena muralis*. Each species is represented by a dedicated BioSample record: SAMN49074188 (*Ae. saxatile*), SAMN49074189 (*C. glauca*), SAMN49074190 (*E. linariifolium*), SAMN49074191 (*N. praecox*; listed in NCBI taxonomy as *Thlaspi praecox*), and SAMN49074192 (*O. muralis*). Scaffold-level genome assemblies are available in GenBank format (FASTA) under the following accession numbers: GCA_051546265.1 (*A. saxatile*), GCA_051988385.1 (*C. glauca*), GCA_054998685.1 (*E. linariifolium*), GCA_051988365.1 (*O. muralis*), and GCA_051546285.1 (*N. praecox*). Each assembly record includes the assembled genome sequence, assembly report files, and standard NCBI metadata. All assemblies are linked to their corresponding BioSample records within the BioProject structure. Raw DNA sequencing data have been deposited in the NCBI Sequence Read Archive (SRA) under the same BioProject and are linked to the respective BioSamples. These datasets include PacBio HiFi long-read sequencing data and Hi-C paired-end sequencing data. All records are publicly released and accessible via the BioProject landing page at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1276498. The genome annotation files, together with original genome assemblies with scaffold names matching the annotation, protein sequences and tables showing external evidence of predicted genes by assembled transcripts or aligned protein sequences are available on GitHub[61]. The orthology and functional annotation of protein coding genes, including additional species, is available in separate GitHub repository[62] or its persistent copy on Zenodo[63]. The functional_annotation directory contains table for each species with *Arabidopsis thaliana* orthologues, GO terms combined from several sources and gene names, descriptions and other

functional annotations from several databases. The detailed description of columns in the tables is in associated README.md file.

Technical Validation

Validation of genome assembly and annotation

Quality control (QC) was performed after each step of the genome assembly pipeline (Fig. 3). Assembly statistics such as N50, L50, and total assembly size were generated using QUAST v5.2.0[64] to assess contiguity and completeness at each stage (Table 6, only the stats from final QC are reported). Genome completeness was evaluated using BUSCO v5.7.1[65] with the Brassicales database (odb10; Table 7). Base-level accuracy and k-mer completeness were assessed using Merqury v1.3[66] (Table 7).

Similarly, as for genome assembly, the annotation completeness was evaluated by BUSCO v5.7.1[65] using Brassicales database (odb10). The score for complete BUSCO genes was 95.9–98.5% across the species, with reduction of only 0.9–1.6 percent points compared to the assembly BUSCO scores (Supplementary Table 5, Table 7).

The reliability of the newly predicted protein-coding gene models was assessed by overlapping them with (1) transcripts assembled from RNA-seq used for annotation and with (2) protein sequences from three well-annotated Brassicaceae species (*Arabidopsis thaliana*, *Arabidopsis lyrata* and *Brassica rapa*) aligned to each newly assembled genome. We acknowledge that this method cannot confirm every real gene model as we used RNA-seq data only from a subset of tissues and from one condition. Also, genes of closer species would have higher probability of being confirmed by aligned proteins. However, this information allows the users of the annotation to divide gene models into several categories of reliability based on the amount of external evidence.

We demonstrate this on distribution of protein lengths of predicted gene models (Fig. 6). Protein lengths distribution of all genes (including less reliable) shows that there are considerable differences between species in the range of short proteins (50–150 amino acid residues), suggesting that for some species (especially *Ae. saxatile*) the short proteins might be over-predicted (Fig. 6a). However, removing gene models without any external evidence results in a protein length spectrum almost identical for all the species used (Fig. 6b). The tables with external evidence for each gene are provided with genome annotations[61] and functional annotations of genes[62]. All the predicted gene models were provided in the genome annotation files to enable the users further filter for the relevant subsets.

Table 6
Assembly statistics for five species (*Aethionema saxatile*, *Cardamine glauca*, *Erysimum linariifolium*, *Noccaea praecox* and *Odontarrhena muralis*) generated using QUAST.

Contig numbers		***Ae. saxatile***	***C. glauca***	***E. linariifolium***	***N. praecox***	***O. muralis***
	Primary	**4267**	**821**	**1910**	**1773**	**822**
	After filtration	**459**	**102**	**85**	**64**	**90**
	Polished	**181**	**53**	**69**	**49**	**57**
Largest contig (Mbp)		28.15	39.23	36.18	38.21	32.47
Total length (Mbp)	Primary	695.77	336.39	260.85	335.90	280.43
	After filtration	287.61	302.78	190.38	254.04	249.17
	Polished	274.50	301.01	189.87	253.53	247.78
GC (%)		36.29	37.85	36.43	39.38	37.60
N50 (Mbp)		20.91	34.78	23.58	33.89	29.99
N75 (Mbp)		19.79	30.58	20.91	27.31	28.12
L50		6	4	4	4	4
L75		10	7	6	6	7

Table 7
Assembly quality assessment for five species (*Aethionema saxatile*, *Cardamine glauca*, *Erysimum linariifolium*, *Noccaea praecox* and *Odontarrhena muralis*) showing BUSCO completeness scores and Merqury k-mer based quality metrics.

		Ae. saxatile	*C. glauca*	*E. linariifolium*	*N. praecox*	*O. muralis*
BUSCO	Complete (%)	95.8	98.2	99.1	98.5	97.4
	Single-copy (%)	91.4	93.3	95.1	94.1	92.6
	Duplicated (%)	4.4	4.9	4	4.4	4.8
	Fragmented (%)	0.4	0.3	0.1	0.2	0.3
	Missing (%)	3.8	1.5	0.8	1.3	2.3
Merqury	qv	62.82	67.46	66.42	67.21	65.88
	Completeness (%)	96.67	78.15	80.67	79.54	85.23

Data Availability

All genome assemblies and raw sequencing data generated in this study are publicly available at NCBI under BioProject accession PRJNA1276498. Genome assemblies are accessible through GenBank under accessions GCA_051546265.1 (*Aethionema saxatile*), GCA_051988385.1 (*Cardamine glauca*), GCA_054998685.1 (*E. linariifolium*), GCA_051988365.1 (*Odontarrhena muralis*), and GCA_051546285.1 (*Noccaea praecox*). Raw PacBio HiFi and Hi-C sequencing datasets are available in the Sequence Read Archive (SRA) and are linked to BioSamples SAMN49074188–SAMN49074192. All datasets are fully released and can be accessed via https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1276498. Genome annotations and associated files are available on GitHub[61]. The orthology and functional annotation of protein coding genes, including additional species, is available in separate GitHub repository[62] or its persistent copy on Zenodo[63].

Code Availability

All data processing and bioinformatics analyses were conducted using publicly available software, following the protocols and manuals provided by each respective tool. Custom scripts and pipelines developed for genome assembly, annotation and orthology are available at the authors' GitHub repositories[61–63,67].

Declarations

Funding

This work was supported by the Czech Science Foundation (projects no. 23-07204M led by F.K. and 26-20931S led by T.M.). Computational resources were provided by the e-INFRA CZ project (ID: 90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. The work was also supported by the project TowArds Next GENeration Crops (CZ.02.01.01/00/ 22_008/0004581) of the ERDF Programme Johannes Amos Comenius.

Author Contribution

F.K. and K.S. designed research; V.B., T.K., F.K., K.J., T.M., S.Š., and P.G.D. arranged and performed sampling; G.Š., J.N. and T.K. performed laboratory work; T.M. and M.A.L. performed karyological analyses; B.H., K.S., M.A.L and F.K. contributed new reagents/analytic tools; M.N.C. and M.D. analyzed data with contribution and guidance from R.W. and L.Y.; M.N.C, M.D. and F.K. wrote the paper with contributions by all co-authors.

Acknowledgement

We acknowledge Hellenic Ministry of the Environment and Energy, General Directorate of Forests and Forest Environment, Directorate of Forest Protection (protocol number: ΥΠΕΝ/ΔΠΔ/16742/1261/19.02.2024) and the Ministry of Environmental Protection of Serbia (permit No 001332782 2024 14850 004 003 501 086) for issuing relevant sampling and research permits. We thank Dmitar Lakušić and Nevena Kuzmanović for helping with locating some of the samples.

Data Availability

All genome assemblies and raw sequencing data generated in this study are publicly available at NCBI under BioProject accession PRJNA1276498. Genome assemblies are accessible through GenBank under accessions GCA_051546265.1 (*Aethionema saxatile*), GCA_051988385.1 (*Cardamine glauca*), GCA_054998685.1 (*E. linariifolium*), GCA_051988365.1 (*Odontarrhena muralis*), and GCA_051546285.1 (*Noccaea praecox*). Raw PacBio HiFi and Hi-C sequencing datasets are available in the Sequence Read Archive (SRA) and are linked to BioSamples SAMN49074188–SAMN49074192. All datasets are fully released and can be accessed via [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1276498] (https:/www.ncbi.nlm.nih.gov/bioproject/PRJNA1276498) . Genome annotations and associated files are available on GitHub 61 . The orthology and functional annotation of protein coding genes, including additional species, is available in separate GitHub repository 62 or its persistent copy on Zenodo 63 .

References

1. Nikolov, L. A. *et al.* Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytol.* 222, 1638–1651 (2019).
2. Hendriks, K. P. *et al.* Global Brassicaceae phylogeny based on filtering of 1,000-gene dataset. *Curr. Biol.* 33, 4052–4068.e6 (2023).
3. German, D. A. *et al.* An updated classification of the Brassicaceae (Cruciferae). *PhytoKeys* 220, 127–144 (2023).
4. Kiefer, M. *et al.* BrassiBase: Introduction to a Novel Knowledge Database on Brassicaceae Evolution. *Plant Cell Physiol.* 55, e3 (2014).
5. Hohmann, N., Wolf, E. M., Lysak, M. A. & Koch, M. A. A Time-Calibrated Road Map of Brassicaceae Species Radiation and Evolutionary History. *Plant Cell* 27, 2770–2784 (2015).
6. Franzke, A., Lysak, M. A., Al-Shehbaz, I. A., Koch, M. A. & Mummenhoff, K. Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends Plant Sci.* 16, 108–116 (2011).
7. Afsharyan, N. P., Golicz, A. A. & Snowdon, R. J. Pangenomic structural variant patterns reflect evolutionary diversification in Brassica napus. 2024.12.19.629214 Preprint at https://doi.org/10.1101/2024.12.19.629214 (2024).
8. Kazakou, E., Dimitrakopoulos, P. G., Baker, A. J. M., Reeves, R. D. & Troumbis, A. Y. Hypotheses, mechanisms and trade-offs of tolerance and adaptation to serpentine soils: from species to

ecosystem level. *Biol. Rev.* 83, 495–508 (2008).

9. Brady, K. U., Kruckeberg, A. R. & Bradshaw Jr., H. D. Evolutionary Ecology of Plant Adaptation to Serpentine Soils. *Annu. Rev. Ecol. Evol. Syst.* 36, 243–266 (2005).
10. Konečná, V., Yant, L. & Kolář, F. The Evolutionary Genomics of Serpentine Adaptation. *Front. Plant Sci.* 11, (2020).
11. Jakovljevic, K. *et al.* Hyperaccumulator plant discoveries in the Balkans: Accumulation, distribution, and practical applications. *Bot. Serbica* 46, 161–178 (2022).
12. Krämer, U. Metal Hyperaccumulation in Plants. *Annu. Rev. Plant Biol.* 61, 517–534 (2010).
13. Španiel, S. & Rešetnik, I. Plant phylogeography of the Balkan Peninsula: spatiotemporal patterns and processes. *Plant Syst. Evol.* 308, 38 (2022).
14. Belser, C. *et al.* Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* 4, 879–887 (2018).
15. Mandáková, T. Chromosome Preparation and Fluorescent In Situ Hybridization. *Methods Mol. Biol.* 2987, 159–180 (2026).
16. Doležel, J., Greilhuber, J. & Suda, J. Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* 2, 2233–2244 (2007).
17. Chapter 11 DAPI Staining of Fixed Cells for High-Resolution Flow Cytometry of Nuclear DNA. in *Methods in Cell Biology* vol. 33 105–110 (Academic Press, 1990).
18. Lipnerová, I., Bureš, P., Horová, L. & Šmarda, P. Evolution of genome size in Carex (Cyperaceae) in relation to chromosome number and genomic base composition. *Ann. Bot.* 111, 79–94 (2013).
19. Temsch, E. M., Temsch, W., Ehrendorfer-Schratt, L. & Greilhuber, J. Heavy Metal Pollution, Selection, and Genome Size: The Species of the Žerjav Study Revisited with Flow Cytometry. *J. Bot.* 2010, 596542 (2010).
20. Russo, A. *et al.* Low-Input High-Molecular-Weight DNA Extraction for Long-Read Sequencing From Plants of Diverse Families. *Front. Plant Sci.* 13, (2022).
21. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* 33, 2759–2761 (2017).
22. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* 11, 1432 (2020).
23. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18, 170–175 (2021).
24. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. Preprint at https://doi.org/10.12688/f1000research.12232.1 (2017).
25. Astashyn, A. *et al.* Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol.* 25, 60 (2024).
26. Kallenborn, F. *et al.* GPU-accelerated homology search with MMseqs2. *Nat. Methods* 22, 2024–2027 (2025).

27. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745 (2016).
28. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* 19, 460 (2018).
29. Ma, Z. *et al.* Omni-Captioner: Data Pipeline, Models, and Benchmark for Omni Detailed Perception. Preprint at https://doi.org/10.48550/arXiv.2510.12720 (2025).
30. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* 10, giab008 (2021).
31. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 (2010).
32. Ghurye, J. *et al.* Integrating Hi-C links with assembly graphs for chromosome-scale assembly. 261149 Preprint at https://doi.org/10.1101/261149 (2018).
33. Ghurye, J., Pop, M., Koren, S., Bickhart, D. & Chin, C.-S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* 18, 527 (2017).
34. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* 3, 99–101 (2016).
35. Smit, A. & Hubley, R. RepeatMasker. (2025).
36. Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat. Methods* 19, 705–710 (2022).
37. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* 34, 867–868 (2018).
38. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* 36, 983–987 (2018).
39. Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* 21, 189 (2020).
40. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26 (2011).
41. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* 117, 9451–9457 (2020).
42. Zhou, C. *et al.* Oatk: a de novo assembly tool for complex plant organelle genomes. *Genome Biol.* 26, 235 (2025).
43. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47, W59–W64 (2019).
44. Ruhlman, T. A. & Jansen, R. K. Plastid Genomes of Flowering Plants: Essential Principles. in *Chloroplast Biotechnology* (ed. Maliga, P.) vol. 2317 3–47 (Springer US, New York, NY, 2021).
45. Bočaj, V., Pongrac, P., Fischer, S. & Likar, M. De novo transcriptome assembly of hyperaccumulating Noccaea praecox for gene discovery. *Sci. Data* 10, 856 (2023).
46. Krueger, F. FelixKrueger/TrimGalore. (2025).

47. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915 (2019).
48. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* 34, 769–777 (2024).
49. Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* 51, D445–D451 (2022).
50. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295 (2015).
51. Dainat, J., Hereñú, D. & Pucholt, P. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF. *GFF Format* 10, (2020).
52. Li, H. Protein-to-genome alignment with miniprot. *Bioinforma. Oxf. Engl.* 39, btad014 (2023).
53. Wlodzimierz, P., Hong, M. & Henderson, I. R. TRASH: Tandem Repeat Annotation and Structural Hierarchy. *Bioinformatics* 39, btad308 (2023).
54. Brown, M. R., Manuel Gonzalez de La Rosa, P. & Blaxter, M. tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets. *Bioinformatics* 41, btaf049 (2025).
55. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238 (2019).
56. Cock, P. J. A., Chilton, J. M., Grüning, B., Johnson, J. E. & Soranzo, N. NCBI BLAST+ integrated into Galaxy. *GigaScience* 4, s13742-015-0080–7 (2015).
57. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421 (2009).
58. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240 (2014).
59. Reiser, L. *et al.* The Arabidopsis Information Resource in 2024. *Genetics* 227, iyae027 (2024).
60. Ahmad, S. *et al.* The UniProt website API: facilitating programmatic access to protein knowledge. *Nucleic Acids Res.* 53, W547–W553 (2025).
61. Duchoslav, M. mduchoslav/Genome_annotations_ultramafic_Brassicaceae. (2026).
62. Duchoslav, M. mduchoslav/Brassicaceae_orthology. (2026).
63. Duchoslav, M. mduchoslav/Brassicaceae_orthology: 3.0. Zenodo https://doi.org/10.5281/zenodo.18741924 (2026).
64. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075 (2013).
65. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212 (2015).
66. 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 (2020).
67. mahnaz65. mahnaz65/Genome_assembly_pipeline. (2025).

Figures

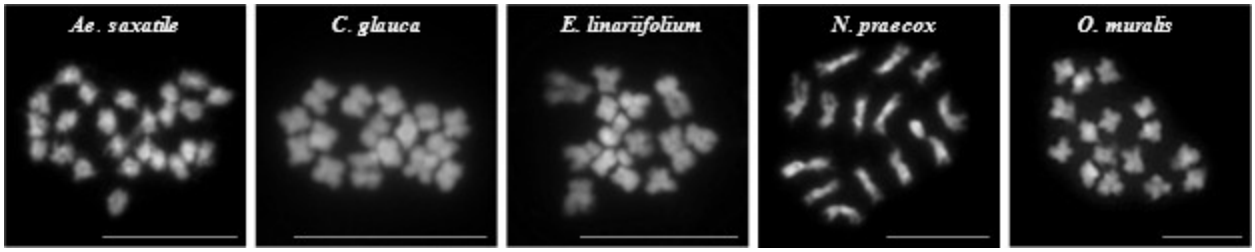

Figure 1

Mitotic chromosomes of *Aethionema saxatile* (2n = 24), *Cardamine glauca* (2n = 16), *Erysimum linariifolium* (2n = 14), *Noccaea praecox* (2n = 14), and *Odontarrhena muralis* (2n = 16) isolated from root tip meristems. Chromosomes were counterstained with DAPI. Scale bar = 10 μm.

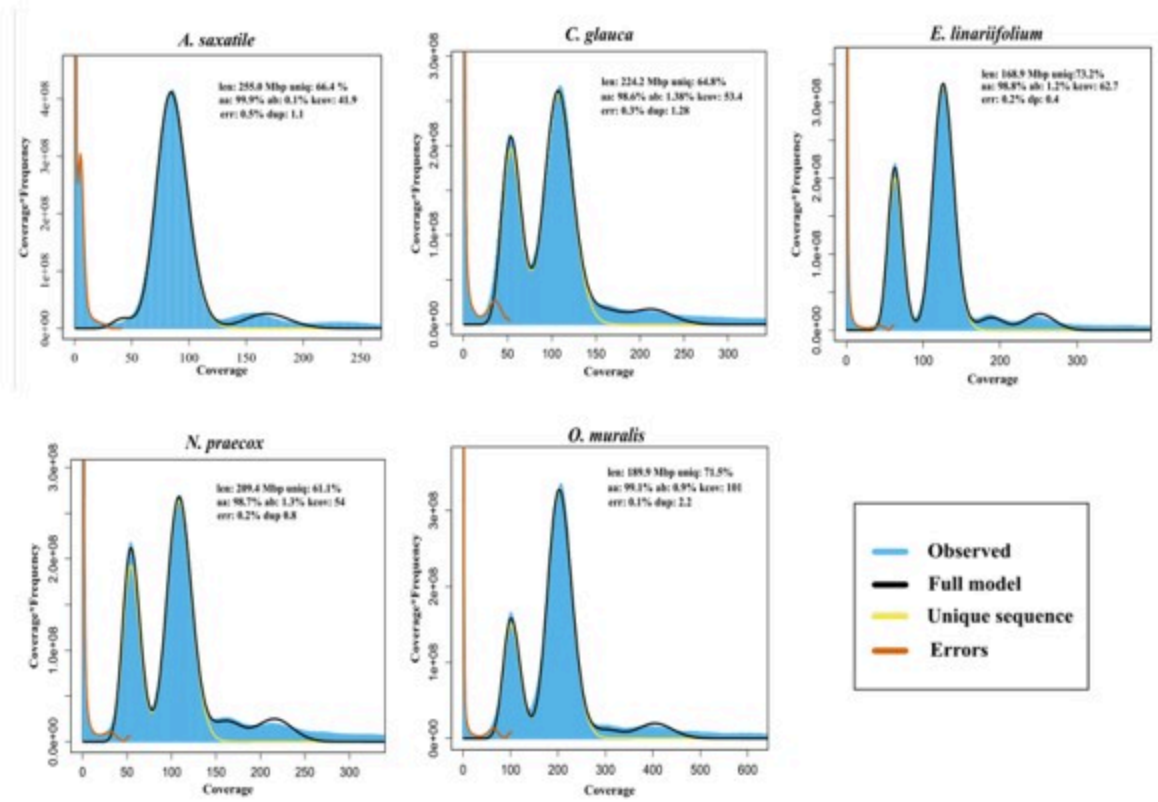

Figure 2

K-mer spectra and genome characteristic estimates for the five Brassicaceae species (*Aethionema saxatile, Cardamine glauca, Erysimum linariifolium, Noccaea praecox* and *Odontarrhena muralis*) inferred from PacBio HiFi reads. Observed k-mer counts (blue), the fitted GenomeScope2 model (black), unique sequence component (yellow) and sequencing error k-mers (orange) are plotted for each species using k =31 and a diploid model (p=2). Model-based estimates of haploid genome size, unique sequence proportion, heterozygosity, error rate, and duplication are indicated within each panel.

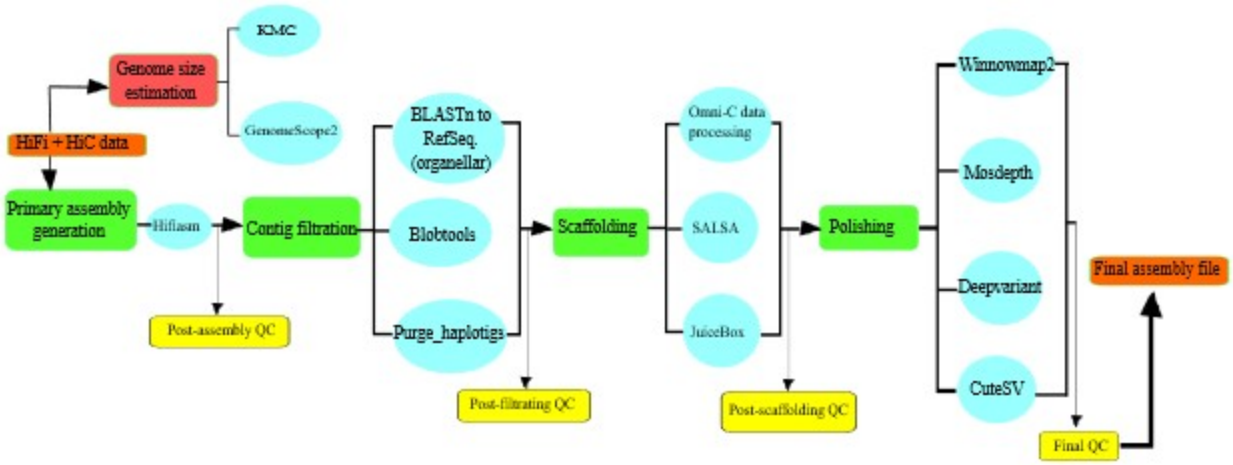

Figure 3

Overview of the genome assembly and quality control pipeline. The diagram summarizes the major phases of the assembly pipeline from genome size estimation to final polishing and quality assessment. Green rectangles indicate the four main assembly phases (primary assembly generation, contig filtering, scaffolding, and polishing). Blue ovals represent software tools used at each phase, whereas yellow rectangles indicate quality control (QC) checkpoints applied after major steps to monitor assembly integrity and accuracy. Genome size estimation is not part of the assembly pipeline itself.

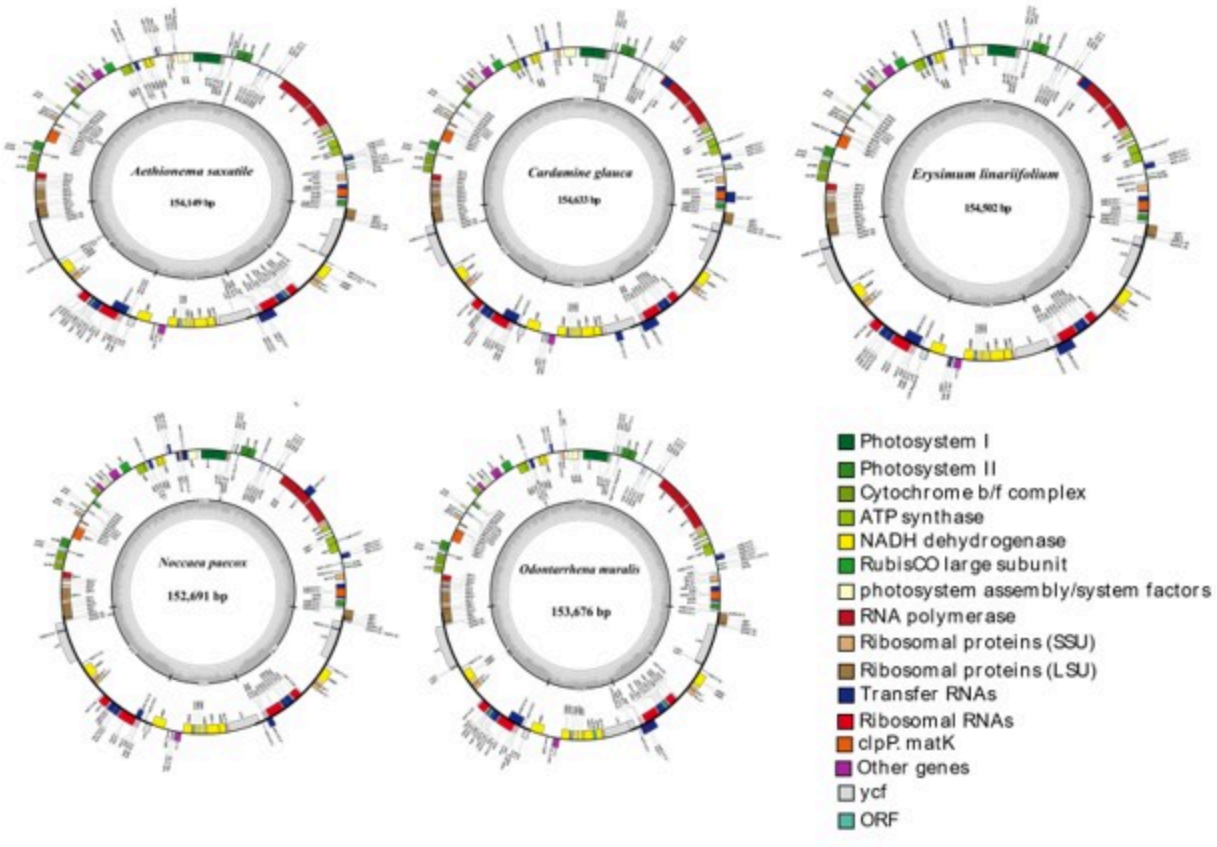

Figure 4

Circular maps of the assembled plastid genomes. Each panel shows one plastid genome with quadripartite structure, including the large single-copy (LSC), small single-copy (SSC), and two inverted repeat regions (IRa and IRb). From the center outward, the inner grey ring represents the plastome backbone with annotated gene density, while the outer ring shows individual plastid genes color-coded by the functional category (legend at right). Gene orientation and genomic positions are shown, and tick marks indicate genome coordinates in kilobases. Circular representations were generated using OGDRAW.

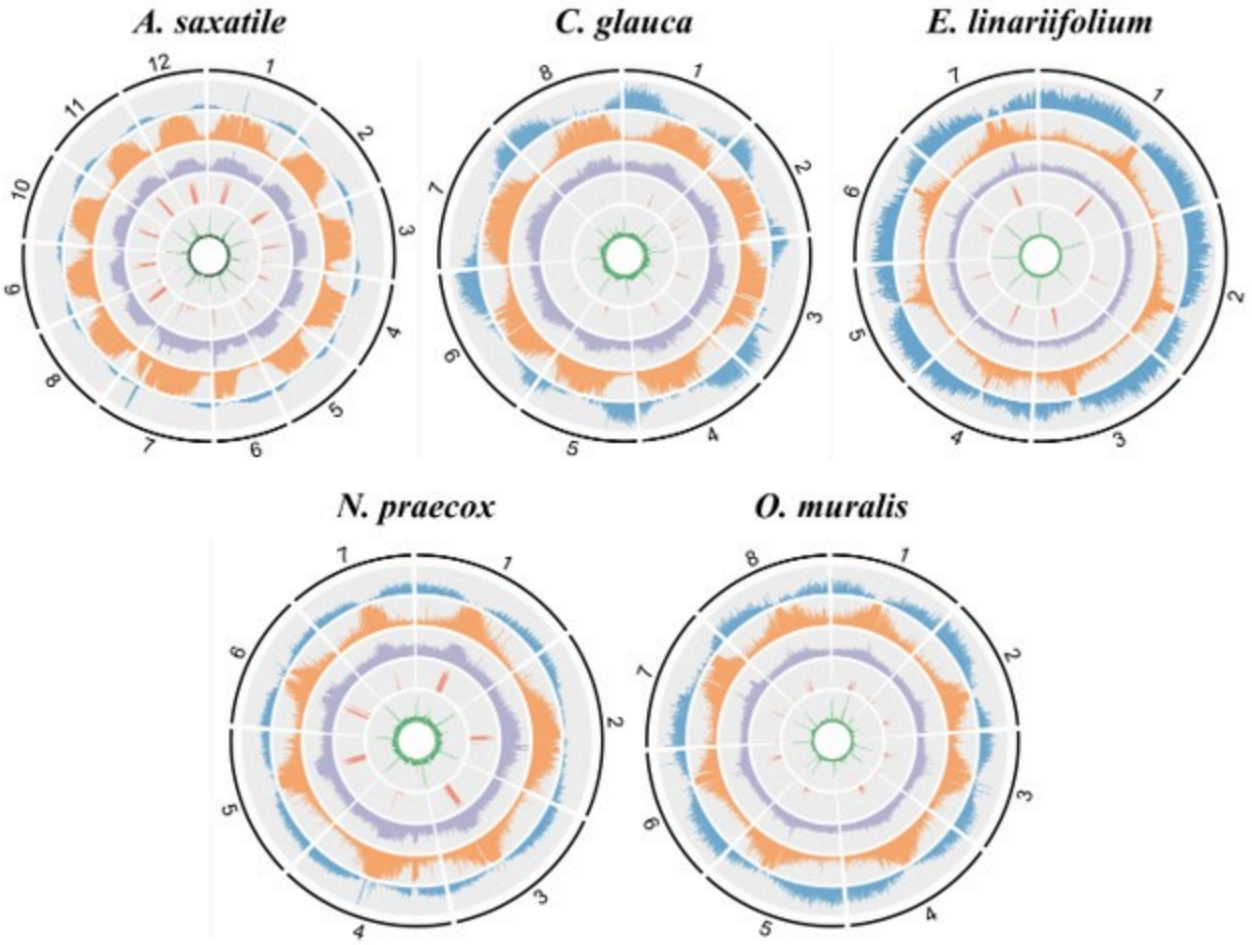

Figure 5

Genome architecture and feature distribution across the new assemblies of the five Brassicaceae species (*Aethionema saxatile*, *Cardamine glauca*, *Erysimum linariifolium*, *Noccaea praecox* and *Odontarrhena muralis*). Circos plots illustrating the distribution of genomic features in 100 kb windows for (A) *Ae. saxatile* (n=12), *C. glauca* (n=8), *E. linariifolium* (n=7), *N. praecox* (n=7), and *O. muralis* (n=8). Each plot displays five concentric tracks from outer to inner: gene density (dark blue, normalized count per 100 kb window scaled 0-100), repeat content (orange, percentage coverage per 100 kb), GC content (purple, percentage per 100 kb with species-specific ranges), centromeric repeat density (red, based on merged TRASH-identified tandem repeat domains), and telomeric repeat density (cyan with dark green outlines, TIDK-identified TTTAGGG motifs normalized to percentage).

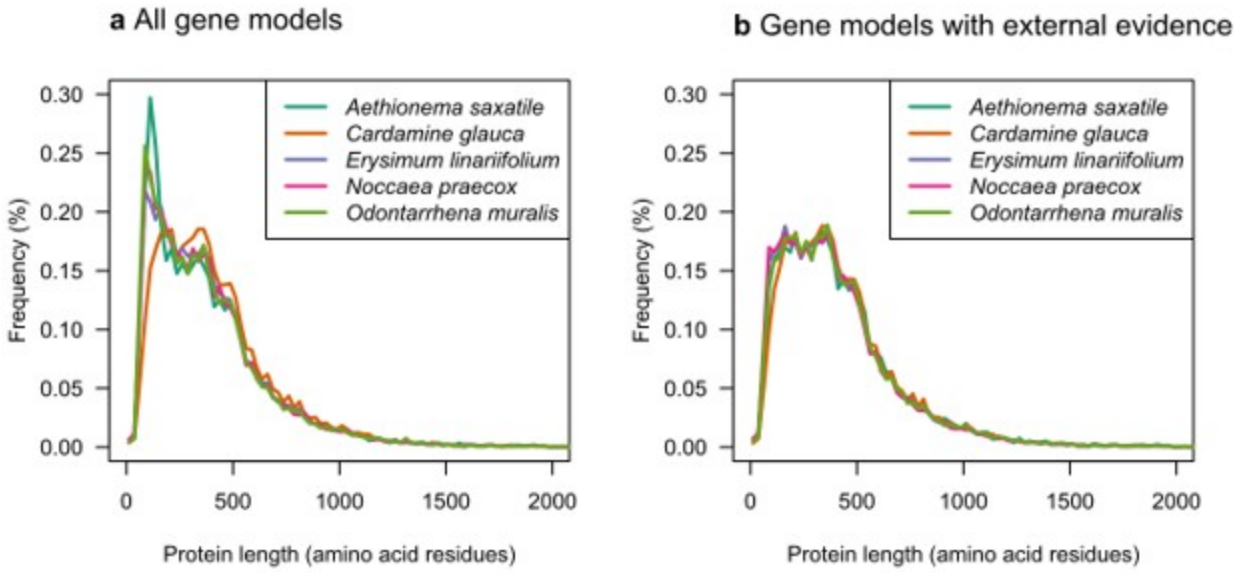

Figure 6

Spectrum of length of predicted protein sequences.(a) All gene models or (b) gene models with some external evidence (overlap with assembled transcript or aligned protein sequence from *Arabidopsis thaliana*, *Arabidopsis lyrata* or *Brassica rapa*) were used for calculation. In case of multiple transcripts per gene, just the longest transcript was used for calculation. The right tail of the plot is cut at 2000 amino acid residues to aid visibility.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable.xlsx