

1 Chromosome-scale genome assembly of *Malcolmia littorea* using long-read sequencing and 2 single-pollen genotyping technologies

3

4 Kenta Shirasawa^{1*}, Kazutoshi Yoshitake², Haruka Kondo³, Shinji Kikuchi³, Keiichiro Koiwai⁴, and
5 Sota Fujii⁵

6

7 1. Department of Frontier Research and Development, Kazusa DNA Research Institute, Japan

8 2. School of Marine Biosciences, Kitasato University, Japan

9 3. Department of Horticulture, Graduate School of Horticulture, Chiba University, Japan

10 4. Laboratory of Genome Science, Tokyo University of Marine Science and Technology, Japan

11 5. Department of Applied Biological Chemistry, Graduate School of Agricultural and Life Sciences,
12 University of Tokyo, Japan

13 *Correspondence: shirasaw@kazusa.or.jp

14

15 Abstract

16 *Malcolmia littorea*, a member of the family Brassicaceae, is adapted to coastal and sandy
17 environments and has become a model in studies of reproductive barriers. However, genomic
18 resources for the species are limited. Here, with the aim of understanding the molecular mechanisms
19 underlying key traits in *M. littorea*, including its survival under harsh conditions, we present a *de*
20 *novo* genome assembly consisting of 10 chromosome-scale sequences. We employed a high-fidelity
21 long-read sequencing technology for genome assembly. To anchor the sequences to chromosomes,
22 we developed a single-pollen genotyping method to construct a genetic linkage map based on SNPs
23 derived from transcriptomes of pollen grains, possessing recombinant haploid genomes. We built a
24 genome assembly consisting of 10 chromosome-scale sequences (214 Mb in total) for *M. littorea*

containing 30,861 predicted genes. A comparative genome analysis and gene prediction indicated that the genome of *M. littorea* is double the size of the *Arabidopsis thaliana* genome, consistent with a whole-genome duplication followed by gene subfunctionalization and/or neofunctionalization in *M. littorea*. This study provides a basis for research on *M. littorea*, an understudied species with ecological and evolutionary significance.

Keywords: Chromosome-scale genome assembly; long-read sequencing; single-pollen genotyping

Introduction

The genus *Malcolmia* (Brassicaceae), a synonym of *Marcus-kochia*¹, contains a diverse group of plants, including *Malcolmia littorea*, which is uniquely adapted to coastal and sandy environments. The species is able to thrive in challenging habitats, often characterized by high salinity, nutrient scarcity, and strong winds. However, specialized genetic mechanisms underlying resilience have not been determined in the species. Despite its ecological significance and potential as a model for understanding plant adaptation to extreme environments, only six articles published up to 2025 were retrieved through a search using the terms “*Malcolmia littorea*” and “*Marcus-kochia littorea*” against the National Center for Biotechnology Information PubMed database (<https://pubmed.ncbi.nlm.nih.gov>), and comprehensive genomic resources are lacking.

In addition to its ecological significance, *M. littorea* has recently been adopted as a model in studies of reproductive barriers. A critical aspect of species delimitation and evolutionary divergence in plants is the presence of reproductive isolation barriers, including interspecific incompatibility. These barriers prevent hybridization between distinct species, thereby maintaining their genetic integrity. In the family Brassicaceae, both self-incompatibility and interspecific incompatibility mechanisms are well-documented, often involving intricate pollen–pistil interactions at the stigma surface, leading to the rejection of "unsuitable" pollen. For instance, studies have

1 identified specific stigmatic proteins, such as Stigmatic Privacy 1 (SPRI1)² and SPRI2³, involved in
 2 the active rejection of pollen from other species in Brassicaceae. *M. littorea* has been used as the
 3 main pollen donor species in these studies; however, comparative genomic analyses are necessary to
 4 further elucidate the molecular mechanisms underlying reproductive barriers. Analyses of the
 5 genetic basis of interspecific incompatibility are crucial for comprehending evolution and
 6 diversification within the family Brassicaceae.

7 Genome sequencing provides an invaluable tool for dissecting the genetic architecture of
 8 adaptive traits, elucidating evolutionary relationships, and guiding conservation strategies.
 9 High-fidelity long-read sequencing, also known as HiFi (PacBio, Menlo Park, CA, USA) sequencing,
 10 is one of the most promising technologies to assemble genome sequences at the chromosome-scale
 11 telomere-to-telomere level, in which a single contig sequence covers an entire chromosome⁴.
 12 However, in some species with complex genome structures, large genome sizes, high heterozygosity
 13 levels, and/or high polyploidy levels, including paleo-ploidy, assembled contig sequences are
 14 chunked into short fragments, insufficient to cover a chromosome. To overcome this issue, the Hi-C
 15 method has been developed to connect short contigs into chromosome-scale sequences⁵.
 16 Alternatively, genetic mapping, based on classical Mendelian principles, enables anchoring genome
 17 sequence fragments to linkage maps to establish chromosome-level sequences⁶. However, genetic
 18 mapping requires mapping populations generated by crossing parental lines. Time and space are
 19 needed to develop and maintain populations and genetically distinct parental lines. Gametophytes
 20 possess haploid genomes containing a single set of chromosomes, which are recombinants of
 21 parental chromosomes. Therefore, gametophytes from a single parent can be used as a mapping
 22 population. Chromosome-scale genome assemblies have been generated by genetic mapping using
 23 gametophytes, including sperms in fish (*Gasterosteus nipponicus*)⁷ and spores in mushroom
 24 (*Lentinula edodes*)⁸.

1 By unraveling the complete genetic blueprint of *M. littorea*, we can gain deeper insights
2 into the genes and pathways responsible for its characteristic stress tolerance and physiological
3 adaptations, its phylogenetic position within the family Brassicaceae, and the molecular
4 underpinnings of its interspecific compatibility or incompatibility with other species. In this study,
5 genome sequencing using a long-read technology, *de novo* assembly via single-pollen genotyping,
6 and comprehensive gene annotation of *M. littorea* were performed, providing a foundational
7 genomic resource for future studies of its biology and evolution.

8

9 **Materials and methods**

10 *Plant materials*

11 *M. littorea* (MAL-LIT-1 [Sp-54]) was provided by Tohoku Univ. Brassica Seed Bank
12 (<https://sites.google.com/dc.tohoku.ac.jp/pbreed/brassica-seed-bank/brassica-seed-bank>). A single
13 plant, Bra27-9, used in previous studies by Fujii et al.^{2,3}, was used in this study.

14

15 *Chromosome observation*

16 Young flower buds of Bra27-9 were fixed with 1:3 acetate: ethanol for at least 2 days and washed
17 with 70% ethanol. Mitotic chromosome slides were prepared with the fixed buds using the
18 enzymatic maceration–squash method⁹. Chromosome images were captured using an OLYMPUS
19 BX-53 fluorescence microscope equipped with a CoolSNAP MYO CCD camera (PhotoMetrics,
20 Huntington Beach, CA, USA) and processed using MetaVue/MetaMorph v7.8 and Adobe Photoshop
21 CS3 v10.0.1. At least five well-spread chromosome slides were used to determine chromosome
22 numbers.

23

24 *Genome size estimation using flow cytometry*

1 Fresh young leaves of approximately $5 \times 5 \text{ mm}^2$ in area were chopped with a razor blade for 20 s
 2 in a Petri dish containing 400 μL extraction buffer (solution A of the Quantum Stain NA UV 2,
 3 Germany) which was added to 1600 μL buffer for DAPI staining (solution B of the Quantum Stain
 4 NA UV2, Germany) and then passed through a nylon sieve (40- μm mesh). After 2 min DAPI
 5 staining, flow cytometric analysis was conducted using CA II cytometer (Partec, Munster, Germany).
 6 *Solanum lycopersicum* was used as the internal standard throughout the analysis. *Arabidopsis*
 7 *thaliana* ecotype Columbia, a member of the same family (Brassicaceae) as *M. littorea*, was used for
 8 genome size comparison. The results were based on the mean of three measurements of material
 9 from different individuals.

10

11 *Genome sequencing and assembly*

12 Genomic DNA was extracted from young leaves using Genomic Tip (Qiagen, Hilden, Germany).
 13 The genomic DNA was sheared to an average fragment size of 30 kb using Megaruptor 2
 14 (Deagenode, Liege, Belgium) in the Large Fragment Hydropore mode. The sheared DNA was used
 15 for HiFi SMRTbell library preparation using the SMRTbell Express Template Prep Kit 2.0 (PacBio).
 16 The resultant library was separated on BluePippin (Sage Science) to remove short DNA fragments
 17 (<15 kb) and sequenced using SMRT Cell 8 M on the Sequel II system (PacBio). The obtained HiFi
 18 reads were assembled using Hifiasm¹⁰ with default parameters. Potential sequences from alternative
 19 alleles were removed using purge_dups¹¹. Organelle genome sequences, identified by sequence
 20 similarity searches of the reported plastid and mitochondrial genome sequences of *Arabidopsis*
 21 *thaliana* using Minimap2¹², were eliminated. Assembly completeness was assessed with
 22 embryophyta_odb10 data using Benchmarking Universal Single-Copy Orthologs (BUSCO)¹³.

23

24 *Pollen protoplast preparation*

1 Pollen grain protoplasts were prepared as described in Fan et al.¹⁴ with minor modifications. Mature
2 pollens were collected from 30 flowers of Bra27-9, washed in pre-buffer solution containing 1/2 MS,
3 1 M Sorbitol, 0.5 M glucose, and 5 mM MES (pH 5.8), filtered through a nylon mesh (diameter = 70
4 μm), and suspended in 15 mL of the pre-buffer. After adding 5 mL of 4 $\times$ enzyme solution (1/2 MS,
5 1.0% Macerozyme R1-0, 2.0% cellulose Onozuka RS, 1 M Sorbitol, 0.5 M glucose, 0.1% BSA, and
6 5 mM MES [pH 5.8]) to the pollen suspension, protoplasts were released at 28°C for 3 h at 350 rpm.
7 The mixture containing pollen protoplasts was centrifuged at 100 $\times g$ for 5 min and pellets were
8 gently resuspended in 10 mL of pre-buffer. This step was repeated twice to remove enzymes and
9 pollen cell wall debris.

10

11 *Cell sorting and transcriptome analysis*

12 The Drop-Seq procedure was used to encapsulate single hemocytes and single mRNA capture beads
13 into fL-scale microdroplets, as described previously.¹⁵ Briefly, a self-built Drop-Seq microfluidic
14 device was prepared by molding polydimethylsiloxane (PDMS; Sylgard 184, Dow Corning Corp.,
15 Midland, MI, USA) in the microchannel structure using a negative photoresist (SU-8 3050, Nippon
16 Kayaku Co., Tokyo, Japan). Using this device, droplets containing a pollen protoplast and a
17 Barcoded Bead SeqB (ChemGenes Corporation, Wilmington, MA, USA) were produced up to 2 mL
18 per sample using a pressure pump system (On-chip Droplet generator, On-chip Biotechnologies Co.,
19 Ltd., Tokyo, Japan). The number of pollen protoplasts was adjusted to 1.2×10^5 cells/buffer.
20 Droplets were collected from the channel outlet into a 50 mL conical tube. Then, droplets were
21 broken promptly and barcoded beads with captured transcriptomes were reverse transcribed with 25
22 μM template switching oligo using Maxima H Minus Reverse Transcriptase (Thermo Fisher
23 Scientific) at 25°C for 30 min, followed by 42°C for 90 min. Then, the beads were treated with
24 Exonuclease I (New England Biolabs, Ipswich, MA, USA) to obtain single-cell transcriptomes

1 attached to microparticles (STAMP). The first-strand cDNAs on beads were amplified using PCR.
2 The beads obtained were distributed throughout PCR tubes (2,000 beads per tube), wherein 1×
3 KAPA HiFi HS Ready Mix (KAPA Biosystems) and 0.8 μM 1st PCR primer were included in a 50
4 μL reaction volume. PCR amplification was achieved using the following program: initial
5 denaturation at 95°C for 3 min; 4 cycles at 98°C for 20 s, 65°C for 45 s, and 72°C for 6 min; 12
6 cycles of 98°C for 20 s, 67°C for 20 s, and 72°C for 6 min; and a final extension at 72°C for 5 min.
7 The amplicons were pooled, double-purified with ×0.6 AMPureXP magnetic beads (Beckman
8 Coulter, Brea, CA, USA), and eluted in 35 μL of ddH₂O. Sequence-ready libraries were prepared
9 using the NexteraXT kit (Illumina, San Diego, CA, USA). A total of 600 pg of each cDNA library
10 was fragmented using a transposome. The amplified library was purified using ×0.6 and ×1.0
11 AMPureXP beads and sequenced (paired-end) on an Illumina NextSeq 500 sequencer, with 20
12 cycles for cell barcodes and UMI as read1 with custom sequence primers and 130 cycles for read2.
13
14 *SNP genotyping, genetic mapping, and chromosome-level sequence construction*
15 Mapping was performed using the STAR aligner¹⁶ via the Drop-seq_alignment.sh script from
16 Drop-seq tools v2.5.1 (<https://github.com/broadinstitute/Drop-seq>). The top 300 cell barcodes with
17 the highest read counts were extracted and saved as individual BAM files. To obtain the
18 ground-truth genotypes, HiFi reads were mapped using minimap2¹² v2.26 with the -ax map-hifi
19 option, and SNP calling was performed using DeepVariant¹⁷ v1.4.0 with the --model_type=PACBIO
20 option. The reads from the top 300 cell barcodes were then used for SNP calling with bbmap
21 (<https://sourceforge.net/projects/bbmap>) through the RNA-seq~SNPcall-bbmap-callvariants script of
22 PortablePipeline v1.4.0 (<https://github.com/c2997108/OpenPortablePipeline>), setting -v "ploidy=2
23 minreads=1". SNPs that were not detected in the HiFi reads were removed as noise. Subsequently, a
24 linkage analysis was performed using the linkage-analysis~SELDLA script of PortablePipeline with

the options -b "-p 0.03 -b 0.03 --NonZeroSampleRate=0.005 --NonZeroPhaseRate=0.1 -r 1000 --RateOfNotNASNP=0.001 --RateOfNotNALD=0.01 --ldseqnum 1 --noNewVcf", enabling chromosome-level genome construction. Finally, manual curation was performed using SELDLA-G v0.9.1 (<https://github.com/c2997108/SELDLA-G>) to build the final genome sequence.

Gene and repetitive sequence prediction

Protein-coding genes were predicted using two approaches, an *ab initio* strategy using Helixer¹⁸ and a homology-based gene prediction strategy using GeMoMa¹⁹. The gene model of *A. thaliana* Araport11²⁰ was employed for homology-based gene prediction. Prediction completeness was assessed with the embryophyta_odb10 data using Benchmarking Universal Single-Copy Orthologs (BUSCO)¹³.

Repetitive sequences in the assembly were identified using RepeatMasker (<https://www.repeatmasker.org>), based on repeat sequences registered in Repbase and a *de novo* repeat library built using RepeatModeler (<https://www.repeatmasker.org>).

Comparative genome structure analysis

Pairwise collinear blocks were detected using MCScanX²¹, with a match score (-k) of 50 and match size (-s) of 50. Results were visualized using a synteny browser, SynVisio²².

Results

*Chromosome number and genome size of *M. littorea**

A total of 20 chromosomes were observed in somatic cells of *M. littorea* buds (Figure 1), suggesting that the chromosome number of *M. littorea* was $2n = 20$. The chromosomes were small and similar in size.

1 The genome size of *M. littorea* was estimated by flow cytometry (Figure 2A) using tomato
2 and *Arabidopsis* as controls. The relative genome sizes of *M. littorea*, *Arabidopsis*, and tomato were
3 0.19, 0.117, and 1.0, respectively. This result indicated that the genome of *M. littorea* was 1.62-fold
4 larger than that of *Arabidopsis* (157 Mb)²³, corresponding to a size of 254 Mb (= 157 Mb × 1.62).
5 Furthermore, 26.1 Gb of HiFi reads, with an N50 length of 19 kb, was obtained and used for a *k*-mer
6 distribution analysis (Figure 2B). The *M. littorea* genome was highly heterozygous (1.8%), and the
7 estimated haploid genome size was 246.1 Mb. Based on these two analyses, the genome size of *M.*
8 *littorea* was estimated to be approximately 250 Mb.

9

10 *Genome sequencing and assembly*

11 The HiFi reads were assembled into 1,316 primary contigs (N50 = 10.0 Mb) spanning a physical
12 distance of 346.1 Mb. The longer primary contig size than the estimated size and the high
13 heterozygosity of the *M. littorea* genome indicated that the primary contigs were a mixture of two
14 haplotype sequences. Therefore, we purged 1,101 contigs (122.2 Mb) as alternative allele sequences.
15 Potential contaminated sequences from organelles were also removed. The final assembly for the *M.*
16 *littorea* genome covered 219.7 Mb, consisting of 118 contigs with an N50 of 13.1 Mb (Table 1). The
17 final assembly was designated as MLI_r1.0. The complete BUSCO score for MLI_r1.0 was 98.0%,
18 of which 92.8% were single-copy BUSCOs (Table 2).

19

20 *Chromosome-level sequence construction by genetic linkage of SNPs from single-pollen*

21 *transcriptome analyses*

22 Protoplasts were prepared from pollen grains harvested from *M. littorea*. Droplets, each of which
23 contained a single protoplast and a Barcoded Bead SeqB (ChemGenes), were generated using a

1 microdevice, as described by Koiwai et al.²⁴ Following the protocol of Koiwai et al.²⁴, single-cell
 2 RNA-Seq libraries were prepared and sequenced to obtain 189.2 million reads.
 3 The sequence reads were split into 3,484,660 subsets in accordance with the barcode sequences
 4 and mapped on the MLI_r1.0 with a mapping rate of 63.1%. Among the 3,484,660 subsets, we
 5 employed the top 300 subsets (29.9% of all data) for genetic analyses. From the read alignment
 6 against MLI_r1.0 as a reference, 330,014 SNPs were detected in the 300 subsets. SNP positions
 7 were validated by a mapping analysis of the HiFi reads back onto MLI_r1.0. A linkage analysis of
 8 these SNPs resulted in a genetic map with 10 linkage groups, referred to as MLI1ch01–MLI1ch10
 9 (Figure 3) consisting of 41,972 SNPs. The total map length was 744.2 cM (Table 3). Twenty-two
 10 MLI_r1.0 contigs (97.2% of the assembly in length) were anchored to the genetic map to establish
 11 10 pseudomolecule sequences as a chromosome-scale genome assembly spanning 213.5 Mb (Tables
 12 1, 3).

13

14 *Predicted genes and repetitive sequences in the M. littorea genome*

15 Protein-coding genes in the *M. littorea* genome were predicted through two approaches. First, using
 16 the *ab initio* strategy of Helixer, 36,880 gene candidates were obtained. After removing
 17 low-confident genes including short exons (1 and 2 bases in length), 26,303 non-redundant genes
 18 were selected as candidates. Second, homology-based gene prediction was carried out using
 19 GeMoMa, in which 27,655 gene models of Araport11 were employed to obtain 27,803
 20 non-redundant gene candidates. Then, the two lists of gene candidates were merged followed by the
 21 removal of redundant genes to obtain 30,861 genes (mean 1,192 bp in length) as a high-confident set
 22 (Table 3). The complete BUSCO score for the 30,861 predicted genes was 97.5% (Table 2).

23 Repetitive sequences occupied a total physical distance of 101.8 Mb (46.3%) in the
 24 MLI_r1.0.genome assembly (219.7 Mb). Nine major types of repeats were identified in varying

1 proportions (Table 4). The dominant repeat types in the chromosome sequences were long-terminal
2 repeats (18.0%) including *gypsy*- (13.0%) and *copia*-type retroelements (4.7%). Repeat sequences
3 unavailable in public databases totaled 42.9 Mb.

4

5 *Comparative analysis of the genome structure of M. littorea*

6 Comparative genome analyses identified 92 pairwise collinear blocks between the genomes of *M.*
7 *littorea* and *A. thaliana* (Supplementary Table S1), ranging from six blocks in MLI1ch03 and
8 MLI1ch06 to 13 blocks in MLI1ch02. Over the genome, two blocks on the *M. littorea* genome
9 corresponded to one block on the *A. thaliana* genome (Figure 4), suggesting a genome-wide
10 duplication in *M. littorea* with respect to the *A. thaliana* genome. Probably due to the genome-wide
11 duplication, 23 pairwise collinear blocks were detected over the *M. littorea* genome (Figure 4,
12 Supplementary Table S2).

13

14 **Discussion**

15 Here, we present a genome assembly for *M. littorea* consisting of 10 chromosome-scale sequences,
16 corresponding to the chromosome number estimated through various approaches in this study
17 (Figure 1) and in a previous report²⁵. This is the first report of a chromosome-scale genome sequence
18 for *M. littorea* and, more broadly, for the genus *Malcolmia*, except for a draft genome assembly
19 (v1.1) of *M. maritima* registered in a database (Brassicales Map Alignment Project, DOE-JGI,
20 <http://bmap.jgi.doe.gov>). Comparative analyses of genome structure revealed that the *M. littorea*
21 genome is doubled with respect to the *A. thaliana* genome (Figure 4), even though the genera are
22 closely related²⁶. This result suggested that a whole-genome duplication event occurred in
23 *Malcolmia* after its divergence from *Arabidopsis*. Alternatively, *Arabidopsis* underwent rapid
24 diploidization. Interestingly, however, probable duplicated genome segments were highly

1 fragmented and not conserved (Figure 4). Furthermore, the single-copy BUSCO scores for *M.*
 2 *littorea* were 92.8% and 91.9% for the genome and gene sequences, respectively (Table 2),
 3 indicating that the gene set in the genome was already diploidized probably via subfunctionalization
 4 and/or neofunctionalization. Therefore, *M. littorea* might be a paleopolyploid. These genetic features
 5 might explain the ecological adaptation of *Malcolmia* to extreme environments.

6 For the chromosome-scale assembly, we employed a genetic mapping strategy. Mapping
 7 populations are required for genetic linkage analyses. F2 populations can usually be used as mapping
 8 populations through self-pollination of an F1 individual derived from a cross between two
 9 homozygous lines as parents. F1 or S1 populations generated from crosses between two
 10 heterozygous lines or through self-pollination of a single heterozygous line, respectively, can also be
 11 used. In this study involving a single line of *M. littorea*, it was impossible to generate such mapping
 12 populations due to the self-incompatibility and lack of multiple, genetically diverse lines as parents.
 13 To overcome this limitation, we used pollens, each possessing a recombinant haploid chromosome
 14 derived from the heterozygous diploid genomes of the parent, as a mapping population.
 15 Single-pollen genotyping has been used in haplotype phasing in genome assemblies in potato²⁷ and
 16 in pear²⁸; however, to the best of our knowledge, this strategy has not been used in genetic mapping,
 17 except in fish⁷ and mushrooms⁸. For genome-wide SNP genotyping-by-sequencing, the
 18 transcriptome was targeted, even though RNA is transcribed from a small part of genomic DNA
 19 (16.8%, calculated from a total gene length of 37.0 Mb divided by the genome length of 219.8 Mb).
 20 The abundance of RNA molecules in cells is advantageous over DNAs (only three copies in a single
 21 pollen consisting of a vegetative nucleus and two sperm cells) for sequencing. Indeed, we
 22 successfully established a genetic map consisting of 41,972 SNPs, which was sufficient to anchor the
 23 genome contigs to the genetic map. Therefore, as shown in this study, single-pollen genotyping
 24 technology, or single-gametophyte genotyping technology, is useful to establish genetic maps even

1 in organisms with large body sizes, long life cycles, or an inability to be farmed as well as in taxa for
2 which it is difficult to prepare mapping populations.

3 The chromosome-scale genome assembly generated in this study provides a crucial
4 resource for determining the genetic basis and mechanisms underlying adaptation to extreme
5 environments, interspecific incompatibility, and evolution and diversification within Brassicaceae.
6 Furthermore, this study represents an initial milestone in uncovering the value of neglected species
7 maintained in seed banks.

8

9 **Acknowledgements**

10 Seeds of *M. littorea* (MAL-LIT-1 [Sp-54]) were provided by the Tohoku Univ. Brassica Seed Bank.
11 We thank Y. Kishida, C. Minami, K. Ozawa, H. Tsuruoka, and A. Watanabe (Kazusa DNA
12 Research Institute) for their technical assistance.

13

14 **Funding**

15 This study was supported in part by JSPS KAKENHI (22H05172, 22H05174, and 22H05181) and
16 the Kazusa DNA Research Institute Foundation.

17

18 **Conflicts of interest**

19 None declared.

20

21 **Data availability**

22 Raw sequence reads and the assembled sequences were deposited in DDBJ (BioProject accession
23 number PRJDB35718). Gene annotation files for coding sequences, putative peptide sequences, and
24 their genome positions are available at Kazusa Genome Atlas (<https://genome.kazusa.or.jp>).

1

2 **References**

- 3 1. Al-Shehbaz, I. A., German, D. A., Mummenhoff, K., and Moazzeni, H. 2014, Systematics,
4 tribal placements, and synopses of the *Malcolmia* S.L. segregates (Brassicaceae). *Harv. Pap.*
5 *Bot.*, **19**, 53–71.
- 6 2. Fujii, S., Tsuchimatsu, T., Kimura, Y., et al. 2019, A stigmatic gene confers interspecies
7 incompatibility in the Brassicaceae. *Nat. Plants*, **5**, 731–41.
- 8 3. Fujii, S., Yamamoto, E., Ito, S., et al. 2023, SHI family transcription factors regulate an
9 interspecific barrier. *Nat. Plants*, **9**, 1862–73.
- 10 4. Aoyagi Blue, Y., Iimura, H., Sato, M. P., and Shirasawa, K. 2025, The impact of
11 telomere-to-telomere genome assembly in the plant pan-genomics era. *Breed. Sci.*, **75**, 3–12.
- 12 5. Dudchenko, O., Batra, S. S., Omer, A. D., et al. 2017, De novo assembly of the *Aedes aegypti*
13 genome using Hi-C yields chromosome-length scaffolds. *Science*, **356**, 92–5.
- 14 6. Tang, H., Zhang, X., Miao, C., et al. 2015, ALLMAPS: robust scaffold ordering based on
15 multiple maps. *Genome Biol.*, **16**, 3.
- 16 7. Yoshitake, K., Ishikawa, A., Yonezawa, R., Kinoshita, S., Kitano, J., and Asakawa, S. 2022,
17 Construction of a chromosome-level Japanese stickleback species genome using ultra-dense
18 linkage analysis with single-cell sperm sequencing. *NAR Genom Bioinform*, **4**, lqac026.
- 19 8. Yoshitake, K., Shirasawa, K., Kojima, K. K., Asakawa, S., Tanaka, N., and Kurokuchi, H. 2025,
20 Nearly telomere-to-telomere genome assembly of the *L. edodes* diploid genome. *Fungal Genet.*
21 *Biol.*, **179**, 104005.
- 22 9. Wang, L., Kikuchi, S., Muto, C., et al. 2015, Reciprocal translocation identified in *Vigna*
23 *angularis* dominates the wild population in East Japan. *J. Plant Res.*, **128**, 653–63.

- 1 10. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. 2021, Haplotype-resolved de
2 novo assembly using phased assembly graphs with hifiasm. *Nat. Methods*, **18**, 170–5.
- 3 11. Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., and Durbin, R. 2020, Identifying
4 and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, **36**,
5 2896–8.
- 6 12. Li, H. 2018, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**,
7 3094–100.
- 8 13. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. 2015,
9 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.
10 *Bioinformatics*, **31**, 3210–2.
- 11 14. Fan, L. M., Wang, Y. F., Wang, H., and Wu, W. H. 2001, In vitro Arabidopsis pollen
12 germination and characterization of the inward potassium currents in Arabidopsis pollen grain
13 protoplasts. *J. Exp. Bot.*, **52**, 1603–14.
- 14 15. Macosko, E. Z., Basu, A., Satija, R., et al. 2015, Highly parallel genome-wide expression
15 profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–14.
- 16 16. Dobin, A., Davis, C. A., Schlesinger, F., et al. 2013, STAR: ultrafast universal RNA-seq aligner.
17 *Bioinformatics*, **29**, 15–21.
- 18 17. Poplin, R., Chang, P.-C., Alexander, D., et al. 2018, A universal SNP and small-indel variant
19 caller using deep neural networks. *Nat. Biotechnol.*, **36**, 983–7.
- 20 18. Stiehler, F., Steinborn, M., Scholz, S., Dey, D., Weber, A. P. M., and Denton, A. K. 2021,
21 Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning.
22 *Bioinformatics*, **36**, 5291–8.
- 23 19. Keilwagen, J., Hartung, F., and Grau, J. 2019, GeMoMa: Homology-based gene prediction
24 utilizing intron position conservation and RNA-seq data. *Methods Mol. Biol.*, **1962**, 161–77.

- 1 20. Cheng, C.-Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., and Town, C. D.
2 2017, Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *Plant*
3 *J.*, **89**, 789–804.
- 4 21. Wang, Y., Tang, H., Debarry, J. D., et al. 2012, MCScanX: a toolkit for detection and
5 evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.*, **40**, e49.
- 6 22. Bandi, V., and Gutwin, C. 2020, Interactive Exploration of Genomic Conservation. *Proc.*
7 *Graph. Interface*, 74–83.
- 8 23. Bennett, M. D., Leitch, I. J., Price, H. J., and Johnston, J. S. 2003, Comparisons with
9 Caenorhabditis (approximately 100 Mb) and Drosophila (approximately 175 Mb) using flow
10 cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus
11 approximately 25% larger than the Arabidopsis genome initiative estimate of approximately
12 125 Mb. *Ann. Bot.*, **91**, 547–57.
- 13 24. Koiwai, K., Koyama, T., Tsuda, S., et al. 2021, Single-cell RNA-seq analysis reveals penaeid
14 shrimp hemocyte subpopulations and cell differentiation process. *Elife*, **10**.
- 15 25. Manton, I. 1932, Introduction to the general cytology of the cruciferae. *Ann. Bot.*, **os-46**,
16 509–56.
- 17 26. Schilbert, H. M., and Glover, B. J. 2022, Analysis of flavonol regulator evolution in the
18 Brassicaceae reveals MYB12, MYB111 and MYB21 duplications and MYB11 and MYB24
19 gene loss. *BMC Genomics*, **23**, 604.
- 20 27. Sun, H., Jiao, W.-B., Krause, K., et al. 2022, Chromosome-scale and haplotype-resolved
21 genome assembly of a tetraploid potato cultivar. *Nat. Genet.*, **54**, 342–8.
- 22 28. Shi, D., Wu, J., Tang, H., et al. 2019, Single-pollen-cell sequencing for gamete-based phased
23 diploid genome assembly in plants. *Genome Res.*, **29**, 1889–99.

24

1 Table 1. *Malcolmia littorea* genome assembly statistics

	MLI_r1.0	MLI_r1.0.pmol
Total size (bp)	219,721,536	213,500,567
No. of sequences	118	10
N50 length (bp)	13,098,664	23,069,650
N90 length (bp)	6,584,281	12,390,297
Gap length (bp)	0	0
No. of genes	n.a.	36,880

2 n.a.: Not analyzed

3

1 Table 2. Completeness of the *Malcolmia littorea* genome assembly and predicted genes

	Genome assembly (%)	Predicted genes (%)
Complete	98.0	97.5
Single-copy complete	92.8	91.9
Duplicated complete	5.2	5.6
Fragmented	0.4	0.2
Missing	1.6	2.3

2

1 Table 3. Summary of linkage groups in *Malcolmia littorea*

Chromosome number	No. of SNPs	Map length (cM)	No. of contigs	Total length (bp)	No. of genes
MLI1ch01	3,289	69.9	3	25,374,001	3,305
MLI1ch02	7,360	72.4	2	27,430,404	4,818
MLI1ch03	2,274	55.5	2	20,095,805	2,211
MLI1ch04	6,536	111.6	4	29,373,502	4,280
MLI1ch05	4,598	119.2	3	23,069,650	3,060
MLI1ch06	3,488	59.2	2	21,105,933	2,099
MLI1ch07	3,462	56.8	2	10,387,743	2,276
MLI1ch08	3,212	49.2	1	12,390,297	2,670
MLI1ch09	3,299	101.5	2	20,485,708	2,433
MLI1ch10	4,454	49	1	23,787,524	2,883
Subtotal (ch01 to ch10)	41,972	744.2	22	213,500,567	30,035
Unassigned contigs	n.a.	n.a.	96	6,340,969	826
Total	41,972	744.2	118	219,841,536	30,861

2 n.a.: Not available

3

1 Table 4. Repetitive sequences in the *Malcolmia littorea* genome

Repeat type	No. of repeats	Total length (bp)	Relative proportion (%)
SINEs	2,715	394,972	0.2
LINEs	8,359	4,715,176	2.1
LTR elements	42,912	39,664,604	18.0
DNA transposons	28,251	9,655,138	4.4
Small RNA	3,197	716,997	0.3
Satellites	4,149	732,066	0.3
Simple repeats	46,962	1,899,466	0.9
Low complexity	10,196	507,682	0.2
Unclassified	115,777	42,933,027	19.5

2

1 **Figure legends**

2 **Figure 1** Chromosomes of *Malcolmia littorea*.

3 Chromosomes observed in a somatic cell of *M. littorea* buds. Scale bar = 5 μ m.

4

5 **Figure 2** Size estimation of the *Malcolmia littorea* genome.

6 **A.** Flow cytometric analysis of *Arabidopsis thaliana* (A), *Solanum lycopersicum* (S), and *Malcolmia*
7 *littorea* (M). **B.** *k*-mer analysis (*k* = 17) with the given multiplicity values.

8

9 **Figure 3** Contact map-style visualization of a single-pollen linkage analysis.

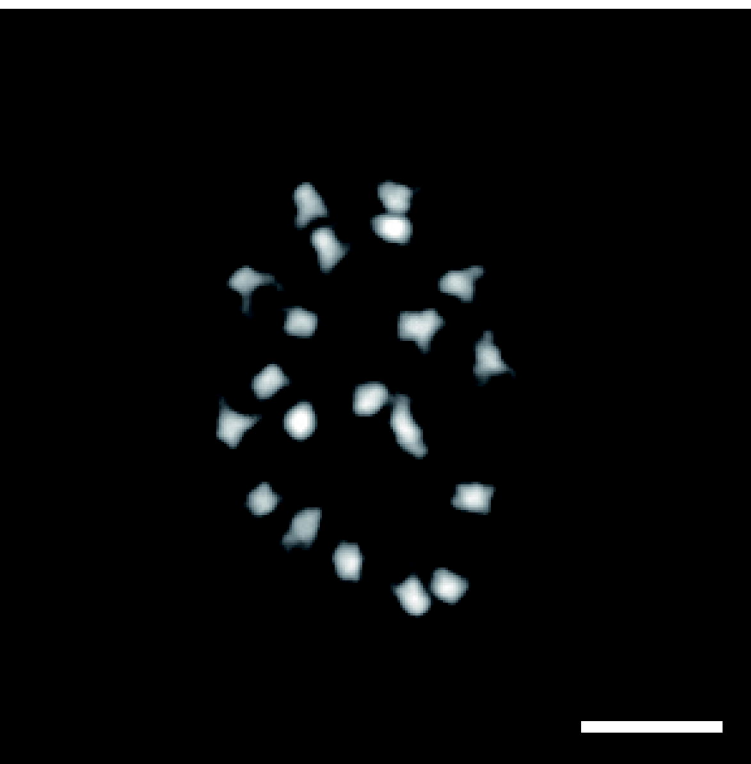
10 Regions in close genetic proximity are shown in red, and distant regions are shown in white.

11

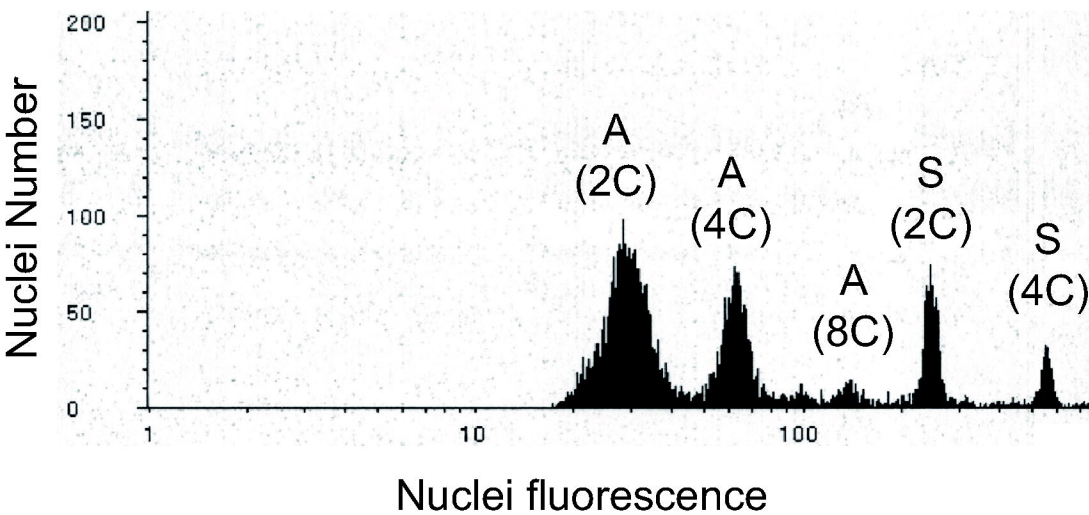
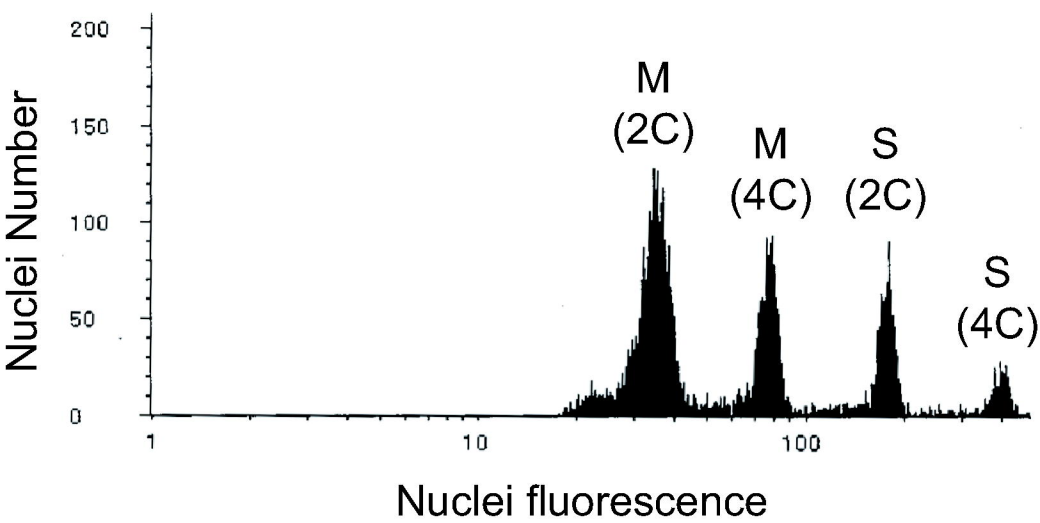
12 **Figure 4** Comparative analysis of genomic structures of *Malcolmia littorea* and *Arabidopsis*
13 *thaliana*.

14 Chromosome structural similarity between *M. littorea* (MLIch01 to MLIch10) to *A. thaliana* (Chr1
15 to Chr5) and within *M. littorea*.

16



A



B

