

Research Article

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

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

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

²School of Marine Biosciences, Kitasato University, Sagami-hara, Kanagawa 252-0373, Japan

³Department of Horticulture, Graduate School of Horticulture, Chiba University, Matsudo, Chiba 271-8510, Japan

⁴Laboratory of Genome Science, Tokyo University of Marine Science and Technology, Minato, Tokyo 108-8477, Japan

⁵Department of Applied Biological Chemistry, Graduate School of Agricultural and Life Sciences, University of Tokyo, Bunkyo, Tokyo 113-8657, Japan

*Corresponding author. Department of Frontier Research and Development, Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan. Email: shirasaw@kazusa.or.jp

Abstract

Malcolmia littorea, a member of the family Brassicaceae, is adapted to coastal and sandy environments and has become a model in studies of reproductive barriers. However, genomic resources for the species are limited. Here, with the aim of understanding the molecular mechanisms underlying key traits in *M. littorea*, we present a *de novo* genome assembly consisting of 10 chromosome-scale sequences. We employed a high-fidelity long-read sequencing technology for genome assembly. To anchor the sequences to chromosomes, we developed a single-pollen genotyping method to construct a genetic linkage map based on SNPs derived from transcriptomes of pollen grains, possessing recombinant haploid genomes. We built a genome assembly consisting of 10 chromosome-scale sequences (215 Mb in total) for *M. littorea* containing 30,266 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.

1. 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. *M. littorea*, so-called sand stock, is able to thrive in challenging habitats, often characterized by high salinity, nutrient scarcity, and strong winds, but has recently declined in abundance by loss of its habitat.^{2,3} 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 sandy coastal environments, 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 identified specific stigmatic proteins, such as Stigmatic Privacy 1 (SPR1)⁴ and SPR2,⁵ involved in the active rejection of pollen from other species in Brassicaceae. *M. littorea*, which possesses self-incompatible system like other species in the Brassicaceae, has been used as the main pollen donor species in these studies; however, comparative genomic analyses are necessary to further elucidate the molecular mechanisms underlying reproductive barriers. Analyses of the genetic basis of interspecific incompatibility are crucial for comprehending evolution and diversification within the family Brassicaceae.

Genome sequencing provides an invaluable tool for dissecting the genetic architecture of adaptive traits, elucidating evolutionary relationships, and guiding conservation strategies. High-fidelity long-read sequencing, also known as HiFi (PacBio, Menlo Park, CA, USA) sequencing, is one of the most promising technologies to assemble genome sequences at the chromosome-scale telomere-to-telomere level, in which a single contig sequence covers an entire chromosome.⁶ However, in some species with complex genome structures,

Received 12 August 2025; revised 16 October 2025; accepted 29 December 2025

© The Author(s) 2026. Published by Oxford University Press on behalf of Kazusa DNA Research Institute.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

large genome sizes, high heterozygosity levels, and/or high polyploidy levels, including paleo-ploidy, assembled contig sequences are chunked into short fragments, insufficient to cover a chromosome. To overcome this issue, the Hi-C method has been developed to connect short contigs into chromosome-scale sequences.⁷ Alternatively, genetic mapping enables anchoring genome sequence fragments to linkage maps to establish chromosome-level sequences.⁸ However, genetic mapping requires mapping populations generated by crossing parental lines. Time and space are needed to develop and maintain populations and genetically distinct parental lines. Gametophytes possess haploid genomes containing a single set of chromosomes, which are recombinants of parental chromosomes. Therefore, gametophytes from a single parent can be used as a mapping population. In conifers, genetic linkage has been analysed using haploid tissues of the giant multicellular megagametophytes from a single maternal plant.⁹ Owing to the great advanced single-cell sequencing technology,¹⁰ this strategy could be applicable to other organisms lacking giant multicellular megagametophytes. Indeed, whereas chromosome-scale genome assemblies have been generated by genetic mapping using gametophytes, including sperms in fish (*Gasterosteus nipponicus*)¹¹ and spores in mushroom (*Lentinula edodes*).¹²

By unravelling the complete genetic blueprint of *M. littorea*, we can gain deeper insights into the genes and pathways responsible for its characteristic stress tolerance and physiological adaptations, its phylogenetic position within the family Brassicaceae, and the molecular underpinnings of its interspecific compatibility or incompatibility with other species. The aim of this study is genome sequencing using a long-read technology, de novo assembly via single-pollen genotyping, and comprehensive gene annotation of *M. littorea*. The genomic resources would provide a foundational genomic resource for future studies of biology and evolution of *M. littorea* and its relatives.

2. Materials and methods

2.1. Plant materials

M. littorea (MAL-LIT-1 [Sp-54]) was provided by Tohoku University Brassica Seed Bank (<https://sites.google.com/dc.tohoku.ac.jp/pbreed/brassica-seed-bank/brassica-seed-bank>)¹³. In accordance with the record in the seed bank, this line was collected from Facultad de Farmacia, Madrid, Spain, in 1965. A single plant, Bra27-9, used in previous studies by Fujii et al.^{4,5} was used in this study.

2.2. Chromosome observation

Young flower buds of Bra27-9 were fixed with 1:3 acetate: ethanol for at least 2 d and washed with 70% ethanol. Mitotic chromosome slides were prepared with the fixed buds using the enzymatic maceration–squash method.¹⁴ Chromosome images were captured using an OLYMPUS BX-53 fluorescence microscope equipped with a CoolSNAP MYO CCD camera (PhotoMetrics, Huntington Beach, CA, USA) and processed using MetaVue/MetaMorph v7.8 and Adobe Photoshop CS3 v10.0.1. At least 5 well-spread chromosome slides were used to determine chromosome numbers.

2.3. Genome size estimation using flow cytometry

Fresh young leaves of approximately 5 × 5 mm² in area were chopped with a razor blade for 20 s in a Petri dish containing

400 µL extraction buffer (solution A of the Quantum Stain NA UV 2, Germany) which was added to 1,600 µL buffer for DAPI staining (solution B of the Quantum Stain NA UV2, Germany) and then passed through a nylon sieve (40-µm mesh). After 2 min DAPI staining, flow cytometric analysis was conducted using CA II cytometer (Partec, Munster, Germany). *Solanum lycopersicum* was used as the internal standard throughout the analysis. *Arabidopsis thaliana* ecotype Columbia, a member of the same family (Brassicaceae) as *M. littorea*, was used for genome size comparison. The results were based on the mean of 3 measurements of material from different individuals.

2.4. Genome sequencing and assembly

Genomic DNA was extracted from young leaves using Genomic Tip (Qiagen, Hilden, Germany). The genomic DNA was sheared to an average fragment size of 30 kb using Megaruptor 2 (Deagenode, Liege, Belgium) in the Large Fragment Hydropore mode. The sheared DNA was used for HiFi SMRTbell library preparation using the SMRTbell Express Template Prep Kit 2.0 (PacBio). The resultant library was separated on BluePippin (Sage Science) to remove short DNA fragments (<15 kb) and sequenced using SMRT Cell 8 M on the Sequel II system (PacBio). The obtained HiFi reads were assembled using Hifiasm¹⁵ with default parameters. Potential sequences from alternative alleles were removed using purge_dups.¹⁶ Organelle genome sequences, identified by sequence similarity searches of the reported plastid and mitochondrial genome sequences of *Arabidopsis thaliana* using Minimap2,¹⁷ were eliminated. In subsequent, potential contaminations from prokaryotes and viruses were also removed or masked using FCS-GX.¹⁸ Assembly completeness was assessed with embryophyta_odb10 data using Benchmarking Universal Single-Copy Orthologs (BUSCO).¹⁹ The genome size and heterozygosity was estimated from *k*-mers (size = 21) of the HiFi reads with Jellyfish²⁰ and GenomeScope 2.0,²¹ respectively.

2.5. Pollen protoplast preparation

Pollen grain protoplasts were prepared as described in Fan et al.²² with minor modifications. Mature pollens were collected from 30 flowers of Bra27-9, washed in pre-buffer solution containing 1/2 MS, 1 M Sorbitol, 0.5 M glucose, and 5 mM MES (pH 5.8), filtered through a nylon mesh (diameter = 70 µm), and suspended in 15 mL of the pre-buffer. After adding 5 mL of 4× enzyme solution (1/2 MS, 1.0% Macerozyme R1-0, 2.0% cellulose Onozuka RS, 1 M Sorbitol, 0.5 M glucose, 0.1% BSA, and 5 mM MES [pH 5.8]) to the pollen suspension, protoplasts were released at 28 °C for 3 h at 350 rpm. The mixture containing pollen protoplasts was centrifuged at 100 × *g* for 5 min and pellets were gently resuspended in 10 mL of pre-buffer. This step was repeated twice to remove enzymes and pollen cell wall debris.

2.6. Cell sorting and transcriptome analysis

The Drop-Seq procedure was used to encapsulate single haemocytes and single mRNA capture beads into fL-scale microdroplets, as described previously.²³ Briefly, a self-built Drop-Seq microfluidic device was prepared by moulding polydimethylsiloxane (Sylgard 184, Dow Corning Corp., Midland, MI, USA) in the microchannel structure using a negative photoresist (SU-8 3050, Nippon Kayaku Co.,

Tokyo, Japan). Using this device, droplets containing a pollen protoplast and a Barcoded Bead SeqB (ChemGenes Corporation, Wilmington, MA, USA) were produced up to 2 mL per sample using a pressure pump system (On-chip Droplet generator, On-chip Biotechnologies Co., Ltd., Tokyo, Japan). The number of pollen protoplasts was adjusted to 1.2×10^5 cells/buffer. Droplets were collected from the channel outlet into a 50 mL conical tube. Then, droplets were broken promptly and barcoded beads with captured transcriptomes were reverse transcribed with 25 μ M template switching oligo using Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific) at 25 °C for 30 min, followed by 42 °C for 90 min. Then, the beads were treated with Exonuclease I (New England Biolabs, Ipswich, MA, USA) to obtain single-cell transcriptomes attached to microparticles. The first-strand cDNAs on beads were amplified using PCR. The beads obtained were distributed throughout PCR tubes (2,000 beads per tube), wherein 1 $\times$ KAPA HiFi HS Ready Mix (KAPA Biosystems) and 0.8 μ M 1st PCR primer were included in a 50 μ L reaction volume. PCR amplification was achieved using the following programme: initial 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 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. The amplicons were pooled, double-purified with $\times 0.6$ AMPureXP magnetic beads (Beckman Coulter, Brea, CA, USA), and eluted in 35 μ L of ddH₂O. Sequence-ready libraries were prepared using the NexteraXT kit (Illumina, San Diego, CA, USA). A total of 600 pg of each cDNA library was fragmented using a transposome. The amplified library was purified using $\times 0.6$ and $\times 1.0$ AMPureXP beads and sequenced (paired-end) on an Illumina NextSeq 500 sequencer, with 20 cycles for cell barcodes and UMI as read1 with custom sequence primers and 130 cycles for read2.

2.7. SNP genotyping, genetic mapping, and chromosome-level sequence construction

SNPs were called from alignments of transcriptome reads on the genome assembly as a reference. Read mapping was performed using the STAR aligner²⁴ via the Drop-seq_alignment.sh script from Drop-seq tools v2.5.1 (<https://github.com/broadinstitute/Drop-seq>). The top 300 cell barcodes with the highest read counts were extracted and saved as individual BAM files. To obtain the ground-truth genotypes, HiFi reads were mapped using minimap2¹⁷ v2.26 with the -ax map-hifi option, and SNP calling was performed using DeepVariant²⁵ v1.4.0 with the -model_type = PACBIO option. The reads from the top 300 cell barcodes were then used for SNP calling with bbmap (<https://sourceforge.net/projects/bbmap>) through the RNA-seq~SNPcall-bbmap-callvariants script of PortablePipeline v1.4.0 (<https://github.com/c2997108/OpenPortablePipeline>), setting -v “ploidy = 2 minreads = 1.” SNPs that were not detected in the HiFi reads were removed as noise. Subsequently, a 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.

2.8. Gene and repetitive sequence prediction

Protein-coding genes were predicted using 2 approaches, an *ab initio* strategy using Helixer²⁶ and a homology-based gene prediction strategy using GeMoMa.²⁷ The gene model of *A. thaliana* Araport11²⁸ was employed for homology-based gene prediction. Prediction completeness was assessed with the embryophyta_odb10 data using BUSCO.¹⁹

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>).

2.9. Comparative genome structure analysis

Pairwise collinear blocks were detected using MCScanX,²⁹ with a match score (-k) of 50 and match size (-s) of 50. Results were visualized using a synteny browser, SynVisio.³⁰

3. Results

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

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

The genome size of *M. littorea* was estimated by flow cytometry (Fig. 2a) using tomato and *Arabidopsis* as controls. The relative genome sizes of *M. littorea*, *Arabidopsis*, and tomato were 0.19, 0.117, and 1.0, respectively. This result indicated that the genome of *M. littorea* was 1.62-fold larger than that of *Arabidopsis* (157 Mb),³¹ corresponding to a size of 254 Mb (= 157 Mb $\times$ 1.62). Furthermore, 26.1 Gb of HiFi reads, with an N50 length of 19 kb, was obtained and used for a *k*-mer distribution analysis (Fig. 2b). The *M. littorea* genome was highly heterozygous (1.8% calculated from the *k*-mer

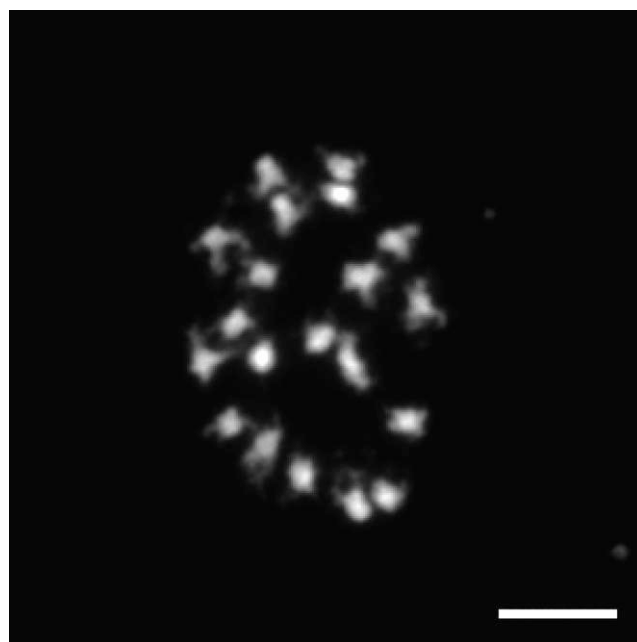


Fig. 1. Chromosomes of *Malcolmia littorea*. Chromosomes observed in a somatic cell of *M. littorea* buds. Scale bar = 5 μ m.

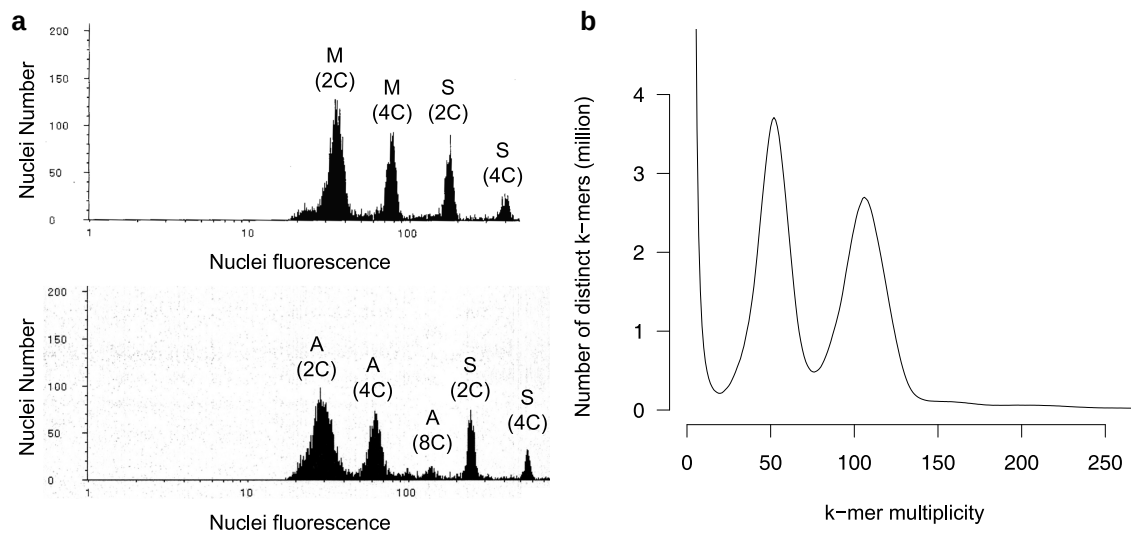


Fig. 2. Size estimation of the *Malcolmia littorea* genome. a) Flow cytometric analysis of *Arabidopsis thaliana* (A), *Solanum lycopersicum* (S), and *Malcolmia littorea* (M). b) k-mer analysis ($k=21$) with the given multiplicity values.

analysis), and the estimated haploid genome size was 246.1 Mb. Based on these 2 analyses, the genome size of *M. littorea* was estimated to be approximately 250 Mb.

3.2. Genome sequencing and assembly

The HiFi reads were assembled into 1,316 primary contigs (N50 = 10.0 Mb) spanning a physical distance of 346.1 Mb. The longer primary contig size than the estimated size and the high heterozygosity of the *M. littorea* genome indicated that the primary contigs were a mixture of 2 haplotype sequences. Therefore, we purged 1,101 contigs (122.2 Mb) as alternative allele sequences. Potential contaminated sequences, 216 contigs (8.9 Mb), from organelles, prokaryotes, and viruses were also removed and that in the primary contigs was masked with 100 Ns. The final assembly for the *M. littorea* genome covered 215.0 Mb, consisting of 29 contigs with an N50 of 13.1 Mb (Table 1). The final assembly was designated as MLI_r1.0. The complete BUSCO score for MLI_r1.0 was 98.0%, of which 92.8% were single-copy BUSCOs (Table 2).

3.3. Chromosome-level sequence construction by genetic linkage of SNPs from single-pollen transcriptome analyses

Droplets, each of which contained a single protoplast and a Barcoded Bead SeqB (ChemGenes), were generated and single-cell RNA-Seq libraries were prepared and sequenced as described by Koiwai et al.³² A total of 189.2 million reads were obtained and split into 3,484,660 subsets in accordance with the barcode sequences and mapped on the MLI_r1.0 with a mapping rate of 63.1%. Among the 3,484,660 subsets, we employed the top 300 subsets (29.9% of all data) for genetic analyses. From the read alignment against MLI_r1.0 as a reference, 330,014 SNPs were detected in the 300 subsets. SNP positions were validated by a mapping analysis of the HiFi reads back onto MLI_r1.0. A linkage analysis of these SNPs resulted in a genetic map with 10 linkage groups, referred to as MLI1ch01–MLI1ch10 (Fig. 3) consisting of 41,972 SNPs. The total map length was 744.2 cM (Table 3). Twenty-two MLI_r1.0 contigs (97.2% of the assembly in length) were

Table 1. *M. littorea* genome assembly statistics.

	MLI_r1.0	MLI_r1.0.pmol
Total size (bp)	215,013,927	213,379,802
No. of sequences	29	10
N50 length (bp)	13,098,664	23,049,850
N90 length (bp)	6,584,281	12,390,297
Gap length (bp)	0	1,300

n.a., not analysed.

Table 2. Completeness of the *M. 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

anchored to the genetic map and connected with 100 Ns to establish 10 pseudomolecule sequences, ie, MLI_r1.0.pmol as a chromosome-scale genome assembly spanning 213.4 Mb (Tables 1, 3). The remaining 7 contigs (total 16 Mb in length) were retained as unplaced sequences of the *M. littorea* genome.

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

Protein-coding genes in the *M. littorea* genome were predicted through 2 approaches. First, using the *ab initio* strategy of Helixer, 31,207 gene candidates were obtained. After removing low-confident genes including short exons (1 and 2 bases in length), 25,839 nonredundant genes were selected as candidates. Second, homology-based gene prediction was carried out using GeMoMa, in which 27,655 gene models of Araport11 were employed to obtain 27,610 nonredundant gene candidates. Then, the 2 lists of gene candidates were

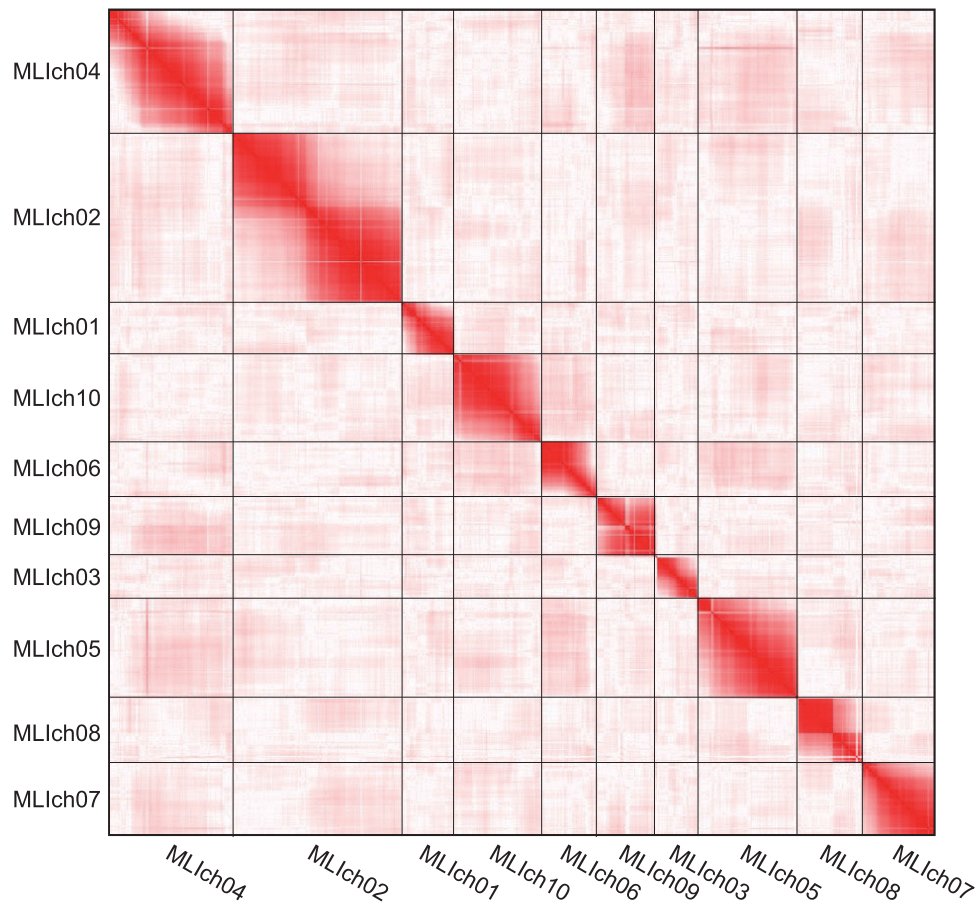


Fig. 3. Contact map-style visualization of a single-pollen linkage analysis. Regions in close genetic proximity are shown in red, and distant regions are shown in white.

Table 3. Summary of linkage groups and pseudomolecule sequences in *M. littorea*.

Chromosome number	No. of SNPs	Map length (cM)	No. of contigs	Total length (bp)	No. of genes
MLI1ch01	3,289	69.9	3	25,352,236	3,305
MLI1ch02	7,360	72.4	2	27,420,504	4,818
MLI1ch03	2,274	55.5	2	20,085,905	2,211
MLI1ch04	6,536	111.6	4	29,343,802	4,280
MLI1ch05	4,598	119.2	3	23,049,850	3,060
MLI1ch06	3,488	59.2	2	21,096,033	2,099
MLI1ch07	3,462	56.8	2	10,377,843	2,276
MLI1ch08	3,212	49.2	1	12,390,297	2,670
MLI1ch09	3,299	101.5	2	20,475,808	2,433
MLI1ch10	4,454	49	1	23,787,524	2,883
Subtotal (ch01 to ch10)	41,972	744.2	22	213,379,802	30,035
Unassigned contigs	n.a.	n.a.	7	1,633,360	231
Total	41,972	744.2	29	215,013,162	30,266

n.a., not available.

merged followed by the removal of redundant genes to obtain 30,266 genes (mean 1,195 bp in length) as a high-confident set (Table 3). The complete BUSCO score for the 30,266 predicted genes was 97.5% (Table 2).

Repetitive sequences occupied a total physical distance of 101.8 Mb (46.3%) in the MLI_r1.0.genome assembly (215.0 Mb). Nine major types of repeats were identified in

Table 4. Repetitive sequences in the *M. 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

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

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

Comparative genome analyses identified 92 pairwise collinear blocks between the genomes of *M. littorea* and *A. thaliana* (Table S1), ranging from 6 blocks in MLI1ch03 and MLI1ch06 to 13 blocks in MLI1ch02. Over the genome, 2 blocks on the *M. littorea* genome corresponded to 1 block on the *A. thaliana* genome (Fig. 4), suggesting a genome-wide

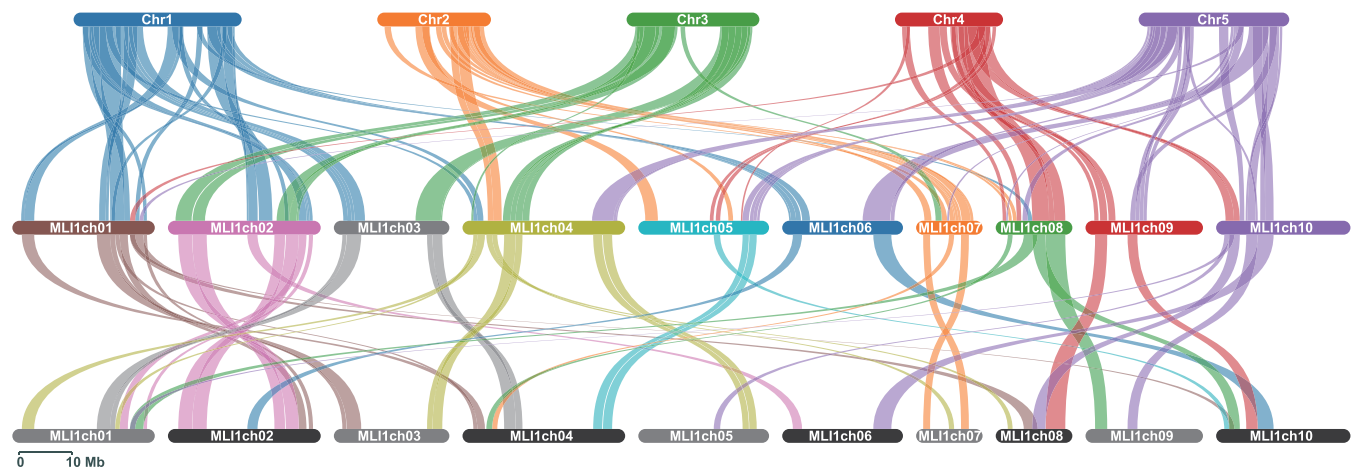


Fig. 4. Comparative analysis of genomic structures of *Malcolmia littorea* and *Arabidopsis thaliana*. Chromosome structural similarity between *M. littorea* (ML1ch01 to ML1ch10) to *A. thaliana* (Chr1 to Chr5) and within *M. littorea*.

duplication in *M. littorea* with respect to the *A. thaliana* genome. Probably due to the genome-wide duplication, 23 pairwise collinear blocks were detected over the *M. littorea* genome (Fig. 4, Table S2).

4. Discussion

Here, we present a genome assembly for *M. littorea* consisting of 10 chromosome-scale sequences, corresponding to the chromosome number estimated through various approaches in this study (Fig. 1) and in a previous report.³³ This is the first report of a chromosome-scale genome sequence for *M. littorea* and, more broadly, for the genus *Malcolmia*, except for a draft genome assembly (v1.1) of *M. maritima* registered in a database (Brassicales Map Alignment Project, DOE-JGI, <http://bmap.jgi.doe.gov>). Comparative analyses of genome structure revealed that the *M. littorea* genome is doubled with respect to the *A. thaliana* genome (Fig. 4), even though the genera are closely related.³⁴ This result suggested that, while *A. thaliana* possesses at least 3 whole-genome duplication (WGD) events, known as α , β , and γ WGDs,³⁵ *M. littorea* might have an additional WGD event or polyploidization. Alternatively, *A. thaliana* underwent rapid diploidization. Interestingly, however, probable duplicated genome segments were highly fragmented and not conserved in the *M. littorea* genome (Fig. 4). Furthermore, the single-copy BUSCO scores for *M. littorea* were 92.8% and 91.9% for the genome and gene sequences, respectively (Table 2), indicating that the gene set in the genome was already diploidized probably via subfunctionalization and/or neofunctionalization. In the Brassicaceae, it is suggested that these genetic features may have conferred higher adaptability and increased tolerance towards adverse environmental conditions.³⁶ The additional WGD event or polyploidization in *M. littorea* might explain its ecological adaptation to sandy coastal environments.

For the chromosome-scale assembly, we employed a genetic mapping strategy. Mapping populations are required for genetic linkage analyses. F_2 populations can usually be used as mapping populations through self-pollination of an F_1 individual derived from a cross between 2 homozygous lines as parents. F_1 or S_1 populations generated from crosses between 2 heterozygous lines or through self-pollination of a single heterozygous line, respectively, can also be used. In this study involving a single line of *M. littorea*, it was impossible to generate

such mapping populations due to the self-incompatibility and lack of multiple, genetically diverse lines as parents. To overcome this limitation, we used pollens, each possessing a recombinant haploid chromosome derived from the heterozygous diploid genomes of the parent, as a mapping population. Megagametophytes can be also used for linkage analysis⁹ but limited to gymnosperms due to tiny size in angiosperms. The great advanced single-cell sequencing technology¹⁰ could change the situation. Single-pollen genotyping has been used in haplotype phasing in genome assemblies in potato³⁷ and in pear³⁸; however, to the best of our knowledge, this strategy has not been used in genetic mapping, except in fish¹¹ and mushrooms.¹² For genome-wide SNP genotyping-by-sequencing, the transcriptome was targeted, even though RNA is transcribed from a small part of genomic DNA (16.8%, calculated from a total gene length of 37.0 Mb divided by the genome length of 219.8 Mb). The abundance of RNA molecules in cells is advantageous over DNAs (only 3 copies in a single-pollen consisting of a vegetative nucleus and 2 sperm cells) for sequencing. Indeed, we successfully established a genetic map consisting of 41,972 SNPs, which was sufficient to anchor the genome contigs to the genetic map. Therefore, as shown in this study, single-pollen genotyping technology, or single-gametophyte genotyping technology, is useful to establish genetic maps even in organisms with large body sizes, long life cycles, or an inability to be farmed as well as in taxa for which it is difficult to prepare mapping populations.

The chromosome-scale genome assembly generated in this study provides a crucial resource for determining the genetic basis and mechanisms underlying adaptation to sandy coastal environments, interspecific incompatibility, and evolution and diversification within Brassicaceae. In addition, the genome resources from this study would improve the conservation genomics of *M. littorea* as a reference for complexity-reduced genome sequencing techniques of population studies.^{2,3} Furthermore, this study represents an initial milestone in uncovering the value of neglected species maintained in seed banks.

Acknowledgements

Seeds of *M. littorea* (MAL-LIT-1 [Sp-54]) were provided by the Tohoku University Brassica Seed Bank. We thank Y. Kishida, C. Minami, K. Ozawa, H. Tsuruoka, and

A. Watanabe (Kazusa DNA Research Institute) for their technical assistance.

Supplementary material

Supplementary data are available at [DNARES](https://dnaresearch.org/) online.

Funding

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

Conflicts of interest

None declared.

Data availability

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

References

- Al-Shehbaz IA, German DA, Mummenhoff K, Moazzeni H. Systematics, tribal placements, and synopses of the *Malcolmia* S.L. segregates (Brassicaceae). *Harv Pap Bot.* 2014;19:53–71. <https://doi.org/10.3100/hpib.v19iss1.2014.n4>
- De Vitis M et al. Integration of genetic and seed fitness data to the conservation of isolated subpopulations of the Mediterranean plant *Malcolmia littorea*. *Plant Biol (Stuttg.)*. 2018;20 Suppl 1:203–213. <https://doi.org/10.1111/plb.12637>
- Del Vecchio S, Giovi E, Izzi CF, Abbate G, Acosta ATR. *Malcolmia littorea*: the isolated Italian population in the European context. *J Nat Conserv.* 2012;20:357–363. <https://doi.org/10.1016/j.jnc.2012.08.001>
- Fujii S et al. A stigmatic gene confers interspecies incompatibility in the Brassicaceae. *Nat Plants.* 2019;5:731–741. <https://doi.org/10.1038/s41477-019-0444-6>
- Fujii S et al. SHI family transcription factors regulate an interspecific barrier. *Nat Plants.* 2023;9:1862–1873. <https://doi.org/10.1038/s41477-023-01535-5>
- Aoyagi Blue Y, Iimura H, Sato MP, Shirasawa K. The impact of telomere-to-telomere genome assembly in the plant pan-genomics era. *Breed Sci.* 2025;75:3–12. <https://doi.org/10.1270/jsbbs.24065>
- Dudchenko O et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science.* 2017;356:92–95. <https://doi.org/10.1126/science.aal3327>
- Tang H et al. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* 2015;16:3. <https://doi.org/10.1186/s13059-014-0573-1>
- Guries RP, Friedman ST, Ledig FT. A megagametophyte analysis of genetic linkage in pitch pine (*Pinus rigida* mill.). *Heredity (Edinb).* 1978;40:309–314. <https://doi.org/10.1038/hdy.1978.31>
- Cuperus JT. Single-cell genomics in plants: current state, future directions, and hurdles to overcome. *Plant Physiol.* 2022;188:749–755. <https://doi.org/10.1093/plphys/kiab478>
- Yoshitake K et al. Construction of a chromosome-level Japanese stickleback species genome using ultra-dense linkage analysis with single-cell sperm sequencing. *NAR Genom Bioinform.* 2022;4:lqac026. <https://doi.org/10.1093/nargab/lqac026>
- Yoshitake K et al. Nearly telomere-to-telomere genome assembly of the *L. edodes* diploid genome. *Fungal Genet Biol.* 2025;179:104005. <https://doi.org/10.1016/j.fgb.2025.104005>
- Tsunoda S, Hinata K, Gomez-Campo C. Preservation of genetic resources. In: Tsunoda S, Hinata K, Gomez-Campo C, editors. *Brassica and wild allies. Biology and breeding.* Japan Scientific Societies Press; 1980. p. 339–354.
- Wang L et al. Reciprocal translocation identified in *Vigna angularis* dominates the wild population in East Japan. *J Plant Res.* 2015;128:653–663. <https://doi.org/10.1007/s10265-015-0720-0>
- Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods.* 2021;18:170–175. <https://doi.org/10.1038/s41592-020-01056-5>
- Guan D et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics.* 2020;36:2896–2898. <https://doi.org/10.1093/bioinformatics/btaa025>
- Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34:3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Astashyn A et al. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol.* 2024;25:60. <https://doi.org/10.1186/s13059-024-03198-7>
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 2015;31:3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* 2011;27:764–770. <https://doi.org/10.1093/bioinformatics/btr011>
- Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun.* 2020;11:1432. <https://doi.org/10.1038/s41467-020-14998-3>
- Fan LM, Wang YF, Wang H, Wu WH. In vitro *Arabidopsis* pollen germination and characterization of the inward potassium currents in *Arabidopsis* pollen grain protoplasts. *J Exp Bot.* 2001;52:1603–1614. <https://doi.org/10.1093/jexbot/52.361.1603>
- Macosko EZ et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell.* 2015;161:1202–1214. <https://doi.org/10.1016/j.cell.2015.05.002>
- Dobin A et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29:15–21. <https://doi.org/10.1093/bioinformatics/bts635>
- Poplin R et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol.* 2018;36:983–987. <https://doi.org/10.1038/nbt.4235>
- Stiehler F et al. Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics.* 2021;36:5291–5298. <https://doi.org/10.1093/bioinformatics/btaa1044>
- Keilwagen J, Hartung F, Grau J. Gemoma: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol. Biol.* 2019;1962:161–177. https://doi.org/10.1007/978-1-4939-9173-0_9
- Cheng C-Y et al. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* 2017;89:789–804. <https://doi.org/10.1111/tpj.13415>
- Wang Y et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 2012;40:e49. <https://doi.org/10.1093/nar/gkr1293>
- Bandi V, Gutwin C. Interactive exploration of genomic conservation. *Proc (Graph Interface).* 2020;74–83. <https://dx.doi.org/10.20380/GI2020.09>
- Bennett MD, Leitch IJ, Price HJ, Johnston JS. Comparisons with *Caenorhabditis* (approximately 100 Mb) and *Drosophila* (approximately 175 Mb) using flow cytometry show genome size in *Arabidopsis* to be approximately 157 Mb and thus approximately

- 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb. *Ann Bot.* 2003;91:547–557. <https://doi.org/10.1093/aob/mcg057>
32. Koiwai K et al. Single-cell RNA-seq analysis reveals penaeid shrimp hemocyte subpopulations and cell differentiation process. *Elife.* 2021;10:e66954. <https://doi.org/10.7554/eLife.66954>
 33. Manton I. Introduction to the general cytology of the cruciferae. *Ann Bot.* 1932;os-46:509–556. <https://doi.org/10.1093/oxfordjournals.aob.a090334>
 34. Schilbert HM, Glover BJ. Analysis of flavonol regulator evolution in the Brassicaceae reveals MYB12, MYB111 and MYB21 duplications and MYB11 and MYB24 gene loss. *BMC Genomics.* 2022;23:604. <https://doi.org/10.1186/s12864-022-08819-8>
 35. Bowers JE, Chapman BA, Rong J, Paterson AH. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature.* 2003;422:433–438. <https://doi.org/10.1038/nature01521>
 36. Kagale S et al. Polyploid evolution of the Brassicaceae during the Cenozoic era. *Plant Cell.* 2014;26:2777–2791. <https://doi.org/10.1105/tpc.114.126391>
 37. Sun H et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat Genet.* 2022;54:342–348. <https://doi.org/10.1038/s41588-022-01015-0>
 38. Shi D et al. Single-pollen-cell sequencing for gamete-based phased diploid genome assembly in plants. *Genome Res.* 2019;29:1889–1899. <https://doi.org/10.1101/gr.251033.119>