Received: 28 June 2020 | Revised: 23 October 2020 | Accepted: 30 October 2020

DOI: 10.1111/1755-0998.13291

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

The chromosome-level genome sequence and karyotypic evolution of *Megadenia pygmaea* (Brassicaceae)

Wenjie Yang[1] | Lei Zhang[1] | Terezie Mandáková[2] | Li Huang[1] | Ting Li[1] | Jiebei Jiang[1] | Yongzhi Yang[3] | Martin A. Lysak[2] | Jianquan Liu[1,3] | Quanjun Hu[1]

[1]Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, China

[2]Central European Institute of Technology (CEITEC), Masaryk University, Brno, Czech Republic

[3]State Key Laboratory of Grassland AgroEcosystem, Institute of Innovation Ecology, Lanzhou University, Lanzhou, China

Correspondence

Quanjun Hu, Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, China.
Email: huquanjun@gmail.com

Funding information

National Natural Science Foundation of China, Grant/Award Number: 41771055 and 31700323; National Key Research and Development program, Grant/Award Number: 2017YFC0505203; Central European Institute of Technology (CEITEC) 2020 project, Grant/Award Number: LQ1601

Abstract

Karyotypic changes in chromosome number and structure are drivers in the divergent evolution of diverse plant species and lineages. This study aimed to reveal the origins of the unique karyotype ($2n = 12$) and phylogenetic relationships of the genus *Megadenia* (Brassicaceae). A high-quality chromosome-scale genome was assembled for *Megadenia pygmaea* using Nanopore long reads and high-throughput chromosome conformation capture (Hi-C). The assembled genome is 215.2 Mb and is anchored on six pseudochromosomes. We annotated a total of 25,607 high-confidence protein-coding genes and corroborated the phylogenetic affinity of *Megadenia* with the Brassicaceae expanded lineage II, containing numerous agricultural crops. We dated the divergence of *Megadenia* from its closest relatives to 27.04 (19.11–36.60) million years ago. A reconstruction of the chromosomal composition of the species was performed based on the de novo assembled genome and comparative chromosome painting analysis. The karyotype structure of *M. pygmaea* is very similar to the previously inferred proto-Calepineae karyotype (PCK; $n = 7$) of the lineage II. However, an end-to-end translocation between two ancestral chromosomes reduced the chromosome number from $n = 7$ to $n = 6$ in *Megadenia*. Our reference genome provides fundamental information for karyotypic evolution and evolutionary study of this genus.

KEYWORDS

Brassicaceae, descending dysploidy, genome assembly, karyotype and chromosome evolution, Qinghai-Tibet Plateau

1 | INTRODUCTION

Karyotypic changes in chromosome number and structure, in addition to polyploidy, are critical drivers in the divergent evolution of diverse plant species and lineages (Stebbins, 1971). Karyotypic changes comprise both chromosome number and large-scale structural changes, which can independently, or in combination, promote evolutionary divergence (Arnegard et al., 2014). The rapid diversification of Brassicaceae arose not only by polyploidy, but through karyotypic changes, providing a useful model system to study the diverse pathways of karyotypic evolution (Lysak et al., 2016; Mandáková & Lysak, 2008). The Brassicaceae is a large angiosperm family comprised of ca. 350 genera and nearly 4,000 species (Kiefer et al., 2014), including scientifically and commercially important species like *Arabidopsis thaliana*, vegetable or oil crops of *Brassica* or *Raphanus*, spices (*Armoracia* and *Eutrema*) and ornamentals (e.g.,

Wenjie Yang and Lei Zhang are contributed equally to this work.

Arabis, *Hesperis*, *Lobularia* and *Matthiola*) (Nikolov et al., 2019). Three major lineages (I, II, and III) or six major clades were identified within the core Brassicaceae (Beilstein et al., 2008; Guo et al., 2017; Huang et al., 2016; Nikolov et al., 2019). The model species *A. thaliana* is included in the lineage I, while the lineage II contains agricultural crops, such as *Brassica napus*, *Brassica rapa* and *Raphanus sativus* (Lv et al., 2020; Nikolov et al., 2019). The number of chromosomes can vary greatly between lineages I and II (Lysak, 2014). Comparative genomics and chromosome painting analyses revealed that the ancestral karyotype of lineage I, the ancestral crucifer karyotype (ACK), comprised eight chromosomes ($n = 8$) and 22 genomic blocks (GBs) (Lysak et al., 2016). The inferred ancestral karyotype of the lineage II, the proto-Calepineae karyotype (PCK: $n = 7$; Mandáková & Lysak, 2008), was found to be derived from the more ancestral PCK genome (ancPCK, $n = 8$) through descending dysploidy, that is chromosome number reduction (Geiser et al., 2016; Mandáková et al., 2018).

Megadenia is a genus of Brassicaceae with a chromosome number $2n = 12$ and relatively few described species, disjunctly distributed across the Qinghai-Tibet Plateau, in northern China, to Asian Russia, and growing at elevation ranges from 400 to 4,000 m above sea level (Artyukova et al., 2014; Dorofeyev, 2004; German & Al-Shehbaz, 2008; Zhou, 2001). All species of *Megadenia* are confined to shady habitats, growing under shrubs and trees or in caves, and have the potential to be horticulturally valuable shade-loving plants (Artyukova et al., 2014). Recent phylogenetic analysis indicated an early divergence of *Megadenia* from other members of lineage II (Guo et al., 2017). The genome sequence of this genus therefore can provide important insights into the karyotype evolution of lineage II shedding light on the earliest karyotypic changes in this clade. In the present study, we report the genome sequence of *M. pygmaea*. It is a small and self-pollinated annual herb with numerous rosette leaves. All flowers on pedicels stretch out of basal rosette leaves. Fruit are indehiscent and valves produce only one seed. This research investigated the detailed chromosome structure of *M. pygmaea* using a chromosome-level de novo genome and chromosome painting analysis. We highlighted the potential mechanism underlying the origin of the six *Megadenia* chromosomes and revealed that an end-to-end chromosome translocation probably mediated the chromosome number reduction from $n = 7$ in the ancestral PCK-like genome to $n = 6$ in the extant *Megadenia* genome. The new reference genome of *M. pygmaea* provides valuable information for advancing the horticultural use of *Megadenia* and aids future investigations into evolution and the uniquely disjunct biogeography of this genus.

2 | MATERIALS AND METHODS

2.1 | Plant material

Young leaves and stems of *M. pygmaea* from two individuals were collected from Ganzi county, Sichuan Province, China (Figure 1a). All fresh materials in the field were immediately frozen and kept in liquid nitrogen until extracting the genomic DNA (gDNA) or total RNA.

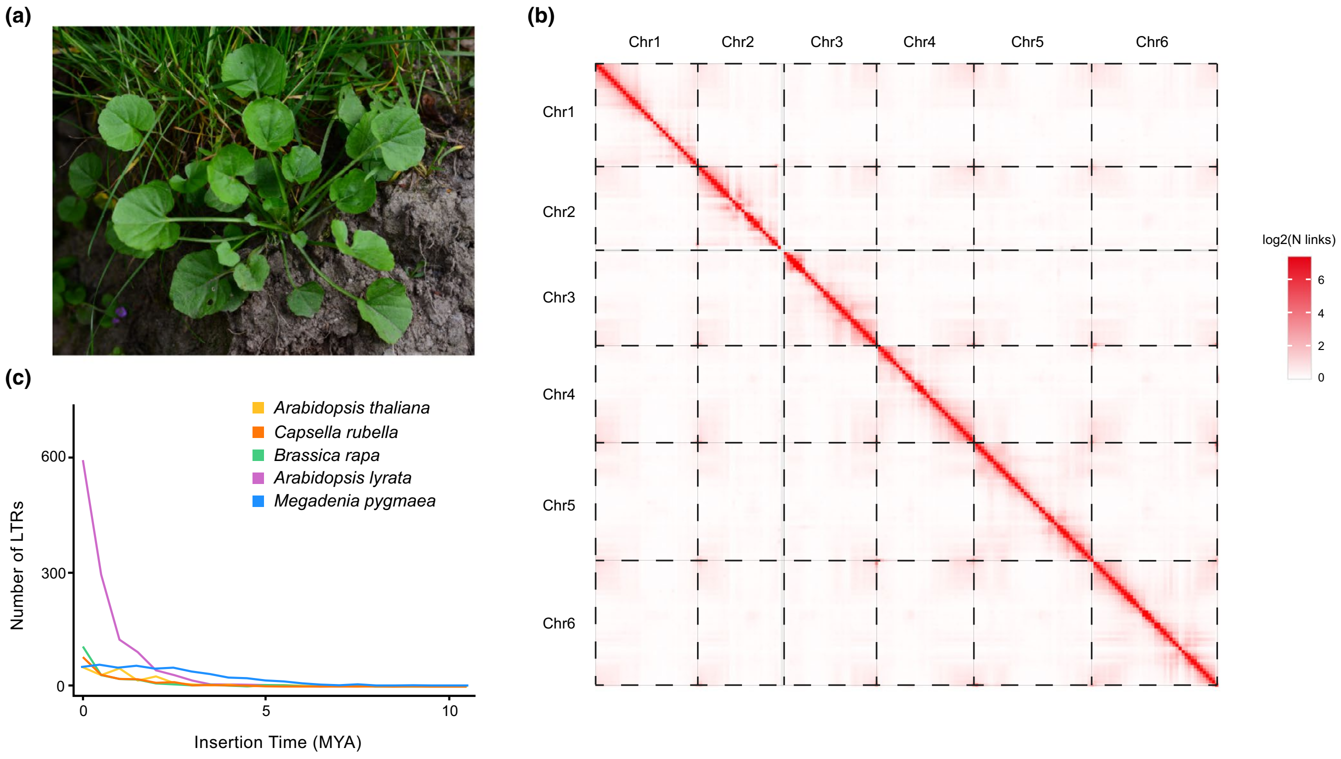

FIGURE 1 (a) Photo of *M. pygmaea*. (b) The Hi-C chromatin interaction map for the six chromosomes of *M. pygmaea*. (c) The evolutionary dynamics of LTR retrotransposons representing intact insertions during the last 10 million years

Leaves from a single individual were used for gDNA extraction and genome assembly. Leaves and stems from the other individuals were used for RNA-seq.

2.2 | Nucleic acid extraction and genome sequencing

High-quality gDNA was extracted using QIAGEN Blood & Cell Culture DNA Kit. We then selected the high molecular weight gDNA (targeting 10–50 kb) using a Blue Pippin (Sage Science, Beverly, MA) and further processed the Nanopore sequencing library with the Ligation sequencing 1D kit (SQK-LSK108, ONT, UK) according to the manufacturer's instructions. We sequenced the resulting library through the GridION X5 sequencer (ONT, UK) at the Genome Center of Nextomics (Wuhan, China). Base calling was further carried out on fast5 files using the ONT Albacore software v0.8.4, and low quality reads (mean_qscore < 7) and adapter sequences were filtered. Sequencing libraries were also prepared with gDNA using Illumina Genomic DNA Sample Preparation Kit and sequenced on an Illumina HiSeq X Ten system in paired-end mode (2 × 150 bp). Unpaired reads, adapter sequences, low-quality reads, and duplicated reads were removed for quality control. The obtained clean data were used for error correction and k-mer analysis. Additionally, total RNAs from the stem and leaf tissues were extracted using Qiagen RNeasy Plant Mini Kits, and sequenced using Illumina HiSeq X Ten system in paired-end mode (2 × 150 bp). The Hi-C library was prepared from 3 g of freshly ground young leaves, using liquid nitrogen with a mortar and pestle. The chromatin extraction, digestion, DNA ligation, purification, and fragmentation were all performed as previously described (van Berkum et al., 2010).

2.3 | Genome assembly

Initial estimates of the genome size were conducted by flow cytometry using *Vigna radiata* for reference (Kang et al., 2014). Genome size was confirmed by k-mer analysis using findGSE v0.1 (Sun et al., 2018) with Illumina short reads. All Nanopore long reads were corrected using canu-correct and trimmed by canu-trim for low-quality bases, and the assembly was performed with Canu v1.7 (Koren et al., 2017). Then, the Hi-C reads were aligned to the assembly using the Juicer v1.6.2 (Durand, Robinson, et al., 2016; Durand, Shamim, et al., 2016). The assembly was scaffolded with Hi-C data using the 3D-DNA v180922 with default parameters (Dudchenko et al., 2017), and manually curated using the Juicebox Assembly Tools v1.11.08 (Dudchenko et al., 2018). The Hi-C scaffolding resulted in six chromosome-level super scaffolds, representing a total of 95.36% of the assembled sequence. We polished the chromosome-level genomes with two iterations using Pilon v1.23 (Walker et al., 2014), and evaluated the completeness of the assembly using BUSCO (Benchmarking Universal Single-Copy Orthologues) v4.1.2 (embryophyte_odb10, 2020-08-05).

2.4 | Evaluation of heterozygosity

We used Illumina sequencing reads to evaluate the level of heterozygosity in *M. pygmaea*. The heterozygosity level was estimated using GenomeScope 2.0 (Ranallo-Benavidez et al., 2020) with 17-mers. The k-mer analysis was performed by Jellyfish v2.29 (Marçais & Kingsford, 2011).

2.5 | Repeats annotation

Repetitive elements in the *M. pygmaea* genome were identified using RepeatMasker v4.0 (Tarailo-Graovac & Chen, 2009) and RepeatModeler v4.07 (Price, Jones, & Pevzner, 2005) with default settings. Intact long terminal repeat (LTR) retrotransposons were identified with LTRharvest v1.5.10 (Ellinghaus et al., 2008) and LTR_Finder v1.06 (Xu & Wang, 2007) with LTR length set to range from 100–5,000 bases and the length between two LTRs set to 1,000–20,000 bases. The LTR_retriever v1.9 (Ou & Jiang, 2018) was used to combine results from LTRharvest and LTR_Finder, and estimate the insertion times of LTR retrotransposon. The insertion times were estimated using $T = K/2\mu$ (Ossowski et al., 2010), where K is the divergence rate and μ is the neutral mutation rate (7×10^{-9} substitutions/site/year).

2.6 | Gene prediction and annotation

A combination of de novo-, homology- and transcript-based methods was used for gene prediction. After quality filtering with Trimmomatic v0.33 (Bolger et al., 2014), a de novo and a genome-guided transcripts assembly was performed on Illumina RNA-seq reads using Trinity v2.6.6 (Haas et al., 2013). Then, transcript-based gene predictions were built with the PASA pipeline v2.1.0 (Haas et al., 2003). Homologues were predicted by mapping protein sequences from *A. thaliana*, *Aethionema arabicum*, *Arabidopsis lyrata*, *B. rapa*, *Capsella rubella*, *Carica papaya*, *Eutrema salsugineum* and *Leavenworthia alabamica* (Table S1) to the *M. pygmaea* genome using exonerate v2.4.0 (Slater & Birney, 2005). A de novo gene prediction was performed with Augustus v3.2.3 with parameters trained using PASA self-trained gene models (Stanke et al., 2004) and with GlimmerHMM v3.0.4 (Majoros et al., 2004). Gene models from the three main sources (i.e., aligned transcripts, de novo predictions and aligned proteins) were merged to produce consensus models by EVidenceModeler v1.1.1 (Haas et al., 2008). The functional annotation for all genes were generated by alignment to public protein databases including Swiss-Prot and TrEMBL (Bairoch & Apweiler, 2000). Protein domains were annotated by searching against InterPro database (Zdobnov & Apweiler, 2001). The GO terms and metabolic pathways were annotated using Blast2GO v2.5 (Conesa et al., 2005) and KEGG databases (Kanehisa et al., 2012). We further extracted collinear paralogous genes and calculated synonymous substitution rates (Ks) to examine potential whole-genome duplication (WGD) events. We used MCScanX (Wang et al., 2012) to detect syntenic blocks (regions with at least five collinear genes) for four species: *A. thaliana*, *B. rapa*, *C. rubella* and *M. pygmaea*. Based on genes in syntenic blocks, we calculated synonymous substitution rates (Ks) to recover the WGD event using codeml in PAML v4.9 (http://abacus.gene.ucl.ac.uk/software/paml.html).

2.7 | Phylogenetic tree construction and divergence time estimation

A phylogenetic tree was built from clusters of gene families for the *M. pygmaea* and several other species representative species of two Brassicaceae lineages (I and II): *A. thaliana*, *A. lyrata*, *Ae. arabicum*, *B. rapa*, *C. rubella*, *E. salsugineum*, *Eutrema yunnanense*, *L. alabamica*, *Raphanus raphanistrum*, *Sisymbrium irio* (Table S1). Gene families were constructed using the OrthoFinder v2.3.12 (Emms & Kelly, 2019) method using all-versus.-all BLASTP alignments (*E*-value ≤ 1e−5). The longest protein encoding sequence at each gene locus for each gene model was retained to remove redundancy caused by alternative splicing. MAFFT v7.313 (Katoh & Standley, 2013) was used to generate sequence alignment for protein sequences in each gene family using the default parameters. For all gene families in the data sets, gene trees were first estimated using FastTree v2.1.11 (Price et al., 2010); these gene trees were then utilized to construct species trees using STAG v.1.0.0 (Emms, 2018). Divergence time was estimated from the phylogenetic tree using MCMCTree from PAML v4.9 (http://abacus.gene.ucl.ac.uk/software/paml.html). Divergence times were determined using a Markov chain Monte Carlo analysis run for 10,000 generations, using a burnin of 1,000 iterations. The calibration time of divergence was obtained from the TimeTree database (Hedges et al., 2006) (http://www.timetree.org/).

2.8 | Gene family expansion and contraction

The expansion or contraction of orthologous gene families was determined using CAFE v4.2 (De Bie et al., 2006). The program uses a birth and death process to model gene gain and loss over phylogenic distance. Gene families that had undergone expansion and/or contraction were calculated using the phylogeny and divergence times with the parameters: *p*-value = 0.05, number of threads = 10.

2.9 | Chromosome preparation

Young inflorescences were fixed in freshly prepared fixative overnight (3:1 ethanol to acetic acid), transferred to 70% ethanol and stored at −20°C. Chromosome spreads were prepared from fixed young flower buds containing immature anthers as previously described (Mandáková & Lysak, 2016b). Chromosome preparations were treated with 100 µg/ml RNase in 2 × sodium saline citrate (SSC; 20 × SSC: 3 M sodium chloride, 300 mM trisodium citrate, pH 7.0) and 0.1 mg/ml pepsin in 0.01 M HCl at 37°C for 60 min and 5 min, respectively. The preparation was then post-fixed in 4% formaldehyde in distilled water and dehydrated by passaging through increasingly pure ethanol (70%, 90% and 100%, 2 min each).

2.10 | Comparative chromosome painting

For comparative chromosome painting (CCP), 674 chromosome-specific BAC clones of *A. thaliana* (The Arabidopsis Information Resource, TAIR; http://www.arabidopsis.org) were used to establish contigs corresponding to the 22 GB and eight chromosomes of the ACK (Lysak et al., 2016). BAC-probes were labelled with biotin-dUTP, digoxigenin-dUTP or Cy3-dUTP by nick translation as previously described (Mandáková & Lysak, 2016a). DNA probes were pooled to follow the given experimental design, ethanol precipitated, dried and dissolved in 20 µl of 50% formamide and 10% dextran sulphate in 2 × SSC. The 20 µl of the dissolved probe was pipetted on a chromosome-containing microscopic slide and immediately denatured on a hot plate at 80°C for 2 min. Hybridization was carried out in a moist chamber at 37°C overnight. Post-hybridization washing was performed in 20% formamide in 2 × SSC at 42°C. Hybridized probes were visualized either as the direct fluorescence of Cy3 or through fluorescently labelled antibodies against biotin and digoxigenin as previously described (Mandáková & Lysak, 2016a). Chromosomes were counterstained with 4´,6-diamidino-2-phenylindole (DAPI, 2 µg/ml) in Vectashield antifade. Fluorescence signals were analysed and photographed using a Zeiss Axioimager epifluorescence microscope equipped with a CoolCube camera (MetaSystems). Images were acquired separately for all four fluorochromes using appropriate excitation and emission filters (AHF Analysentechnik). The four monochromatic images were pseudocolored, merged and cropped using Photoshop CS (Adobe Systems) and ImageJ (National Institutes of Health).

3 | RESULTS

3.1 | Genome assembly and annotation

We generated a total of 51.0 Gb data for genome assembly and gene predictions (Table S2). A total of 17.2 Gb raw data was obtained for long-reads sequencing. After filtering, 13.6 Gb data with a mean reads length of 21.1 kb was recovered. The N50 of reads was 29.9 kb and the longest read was 153.2 kb. The estimated genome size was 219–260 Mb using flow cytometry and k-mer analysis (Figures S1 and S2). The assembled genome is 215.4 Mb in length and the contig N50 is 1.81 Mb. Furthermore, we anchored these contigs into six chromosomes with Hi-C reads using 3D-DNA (Dudchenko et al., 2017). This assembled chromosome-scale genome is 215.2 Mb in length with chromosome N50 = 34.8 Mb (Table 1, Figure 1b). In addition, we used Pilon to polish the genome assembly twice. Genome assembly completeness evaluation suggests a total of 98.9% complete BUSCOs were present (Table S3). The heterozygosity level was estimated to be ~0.4% in the *M. pygmaea* genome (Figure S3, Table S4). The low level

TABLE 1 Overview of the *M. pygmaea* draft genome

Number of pseudo-chromosomes	6
Total length of scaffolds (Mb)	215.2
Super scaffold N50 (Mb)	34.8
Super scaffold N90 (Mb)	27.1
Mean super scaffold length (Mb)	34.1
Contig N50 (Mb)	1.81
Number of genes	25,383
Mean transcript length (bp)	2,628
Mean CDS length (bp)	234
Mean exons per gene	5.4
Mean exon length (bp)	281
Mean intron length (bp)	233
GC content (%)	37.1
Gap content (%)	0.2
Transposable elements (%)	42.6

of heterozygosity and the high-quality genome assembly are largely consistent with the self-pollinated reproductive system and continuous inbreeding of this species. A total of 91.79 Mb (42.66%) of the assembled *M. pygmaea* genome is composed of repetitive sequences (Table 1). Among these repetitive elements, most are LTR retrotransposons, spanning 25.21% of the assembled genome, including 23.93% of intact LTR retrotransposons, followed by DNA transposons (7.03%) and LINEs (2.90%) (Table S5). The insertions of the LTR-RTs in *M. pygmaea* occurred earlier than in *A. lyrata* (Figure 1c). In total, 25,383 genes were predicted, with an average gene length, coding sequence length and an average exon number of 2,628 base pairs (bp), 234 bp and 5.4 exons, respectively (Table 1). The gene prediction showed 95.3% coverage of complete BUSCOs and 70.4% of predicted genes were supported by RNA evidence (Tables S6, S7). In our assembly, 97.88% of the genes (24,846 of 25,383) were annotated on six chromosomes, and only 2.12% (537 of 25,383) remained on unplaced scaffolds. These statistics revealed that the newly assembled genome had high coverage and accuracy in genic regions. Among the 25,383 predicted genes, total 98.14% of the genes were annotated in Swissprot, InterPro, GO and KEGG Pathway databases (Table S8). The *M. pygmaea* genome contains a similar number of transcription factors (TFs) (1,571) as these Brassicaceae species (Table S9; http://www.transcriptionfactor.org).

3.2 | Phylogeny and whole-genome duplication

A total of 336,669 coding sequences from *M. pygmaea* and genomes representing the two Brassicaceae lineages (I and II) were clustered into 43,882 gene families. Species were grouped into phylogenetic lineages according to their COG gene profiles. *M. pygmaea* shared a total of 16,711 with lineage I species and 16,945 with lineage II, with

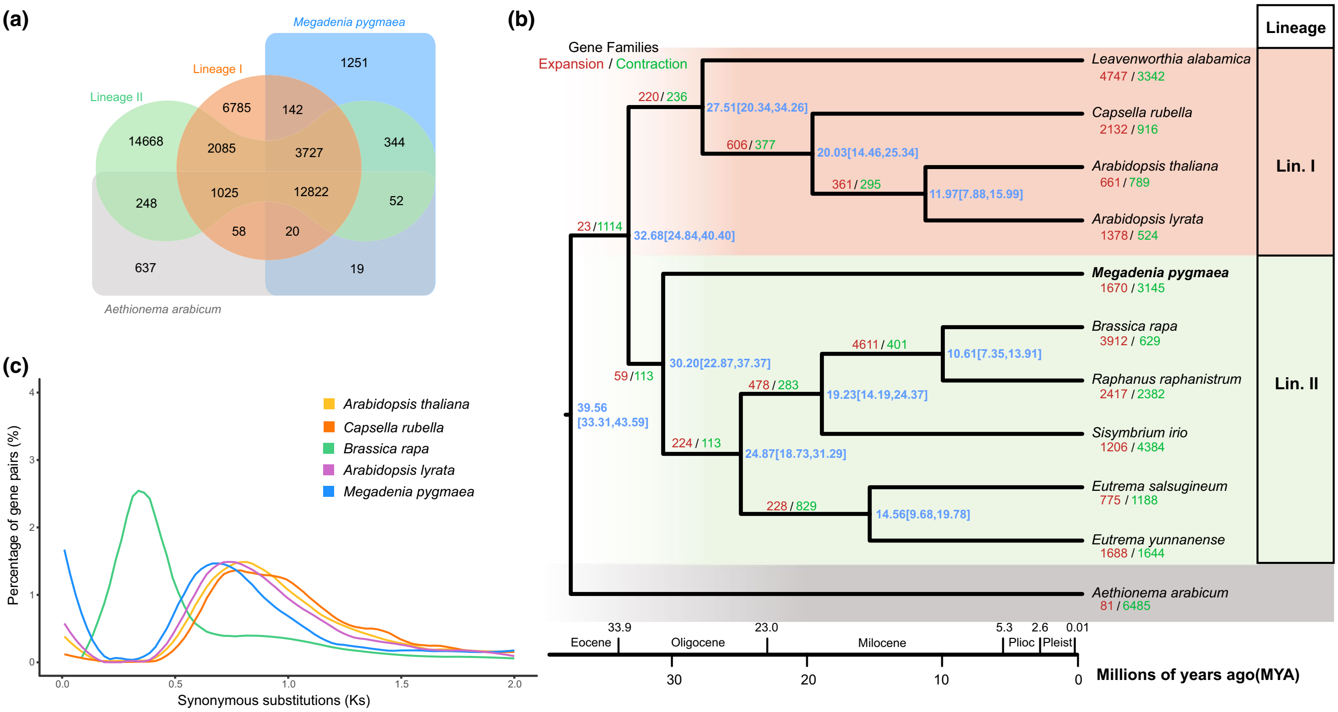

FIGURE 2 (a) Clusters of ortholog groups (COGs) shared between *M. pygmaea* and other Brassicaceae species grouped according to their assignment to phylogenetic Lineages in Brassicaceae (I: *A. thaliana*, *A. lyrata*, *C. rubella* and *L. alabamica*; II: *B. rapa*, *E. salsugineum*, *E. yunnanense*, *R. raphanistrum* and *S. irio*). (b) The Ks values of *M. pygmaea* and other Brassicaceae species. (c) The phylogenetic placement of *M. pygmaea*, divergence time and gene family expansions (red) and contractions (green) displayed on a maximum likelihood tree constructed from 4,245 shared single-copy gene families. The estimated divergence times (in million years ago, blue). Brassicaceae lineage I was represented by *A. thaliana*, *A. lyrata*, *C. rubella* and *L. alabamica*, and lineage II by *B. rapa*, *E. salsugineum*, *E. yunnanense*, *R. raphanistrum* and *S. irio*

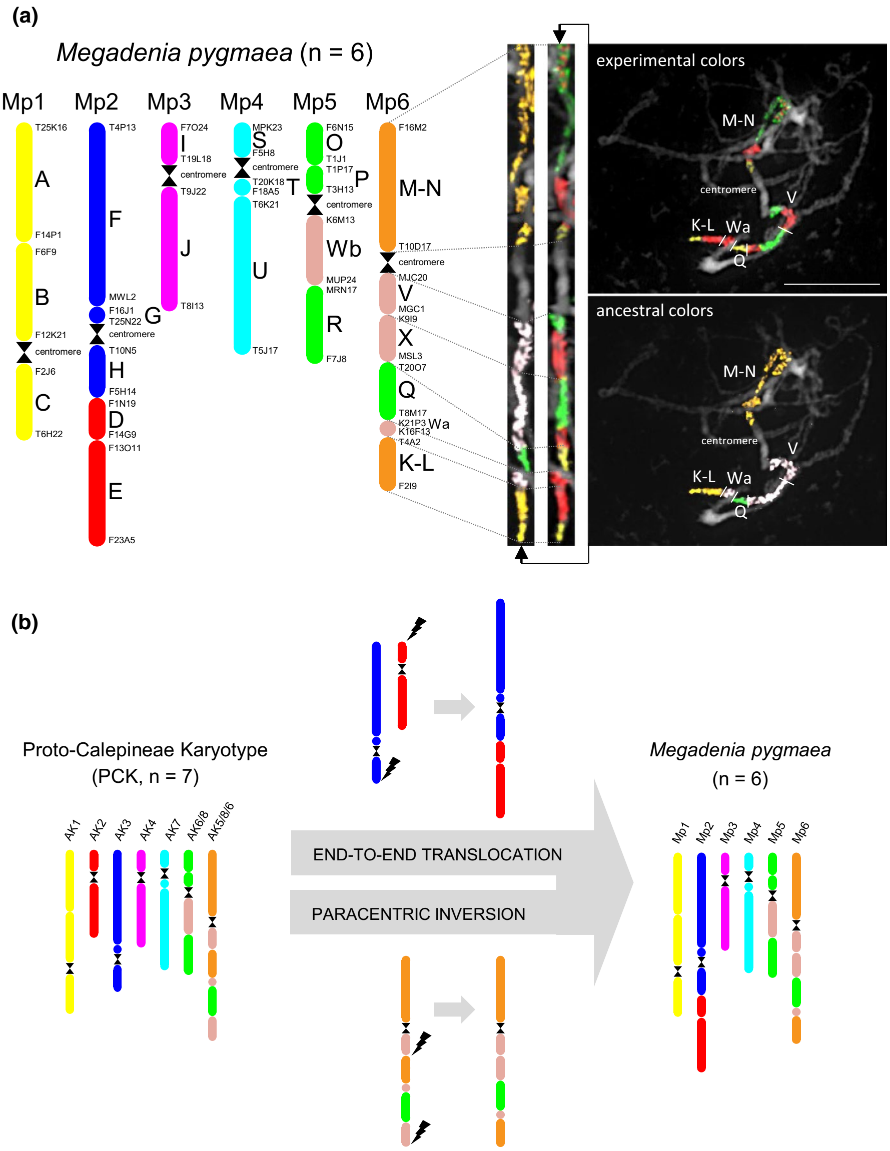

FIGURE 3 (a) Comparative karyotype based on CCP analysis showing the position of 22 genomic blocks (A–X) on six *Megadenia* chromosomes (Mp1-Mp6) and multicolour CCP on pachytene chromosomes of *M. pygmaea* revealing the structure of Mp6. Colour coding reflects the position of genomic blocks on the eight chromosomes in ACK; *A. thaliana* BAC clones delimiting each block are shown. Differentially labelled BAC painting probes on the Mp6 pachytene bivalent are shown in "experimental colours" (red/green/yellow fluorescence) and pseudo-coloured following the colour code of the eight chromosomes of ACK. Chromosomes were counterstained by DAPI. Scale bar, 10 μm. (b) Chromosomal rearrangements illustrating the origin of *Megadenia* genome (*n* = 6) from PCK-like genome (*n* = 7) are displayed. Black lightning symbols indicate chromosomal breakpoints

1,251 gene families unique to *M. pygmaea* (Figure 2a). Whole-genome duplication (WGD) analyses based on collinear paralogous genes revealed that *M. pygmaea*, along with *A. thaliana* and *C. rubella*, did not experience an independent WGD subsequent to the Brassicaceae-specific At-α WGD (Kiefer et al., 2014) (Figure 2b). However, consistent with previous studies, *B. rapa* had a clade-specific whole genome triplication (Cheng et al., 2014; Zhang et al., 2018). This further supports the cytogenetic evidence of the diploid status of *M. pygmaea*. *M. pygmaea* was placed as an independent clade of lineage II, divergent from other representatives in the phylogenetic tree (Figure 2c). *M. pygmaea* was estimated to diverge from other lineage II genera around 30.20 (22.87–37.37) million years ago.

3.3 | Gene expansion/contraction and species-specific genes in *M. pygmaeas*

A total of 54 and 187 gene families significantly ($p < .05$) expanded and contracted in *M. pygmaea*, respectively, of the 1,670 and 3,145 that significantly differed among other lineage II genomes (Figure 2c). The significantly expanded and contracted gene families contain 201 and 244 genes, respectively. The functional annotation of these genes revealed that expanded genes were involved in defense response, regulation of cellular response to stress, response to stimulus, insect, fungus, incompatible interaction and other organism (Table S10). We extracted 1,715 species-specific genes in the *M. pygmaea* genome. These genes were enriched in cellular macromolecule metabolic process, cellular process and DNA replication (Table S11).

3.4 | Comparative chromosomal painting

All painting probes (Lysak et al., 2016; Schranz et al., 2006) each identifying a unique chromosome region confirmed the diploid status of the *Megadenia* genome. The complete comparative chromosomal map of *M. pygmaea* (Figure 3), constructed by CCP, had similarities and notable differences to the structure of ancestral

Brassicaceae genomes: ACK, ancPCK and PCK. Three chromosomes of *M. pygmaea* (Mp1, Mp3 and Mp4) structurally mirrored three ancestral chromosomes (AK1, AK4 and AK7) found in ACK, ancPCK and PCK. Among the three remaining chromosomes, Mp5 was homologous to chromosome AK6/8 (GB association O + P+Wb + R) in ancPCK and PCK. Chromosome Mp6 is homologous to PCK-specific chromosome AK5/8/6 (GBs [M–N], V, X, Q, Wa and [K–L]). However, it contains a 9.92 Mb *Megadenia*-specific paracentric inversion on its bottom (long) arm, with breakpoints between GBs V and (K-L) and the (sub)telomere (Figure 3b). Chromosome Mp2 was formed by an end-to-end translocation (EET) merging ancestral chromosomes AK2 and AK3 (Figure 3b), revealing dysploidy resulting in a reduction from seven to six chromosomes. The presence of the PCK-specific chromosome AK5/8/6 (Mp6) in *M. pygmaea* suggests descent from a seven chromosome-containing ancestral PCK-like genome. We also compared the *M. pygmaea* genome with *A. thaliana* and *C. rubella* genome by MCScanX (Wang et al., 2012) using the same method as published previously (Kang et al., 2020). The syntenic relationships, order and orientation of the 22 GBs by CCP produced the same schematic diagram of the *M. pygmaea* genome (Figures S4 and S5).

4 | DISCUSSION

Our study produced a high-quality genome of a shade-loving plant, *M. pygmaea*, with potential horticultural use. The genome sequence presented here is therefore useful for its domestication and breeding in the future. In addition, the genus *Megadenia* is distributed from the high-altitude Qinghai-Tibet Plateau to the low-altitude northern Russia. The reported reference genome here can be used to decipher such biogeographic connections through resequencing genomes of more populations for this species and other congeneric species across these disjunct distributions.

Another main aim in the present study is to use this genome of this species to clarify the karyotype evolution in lineage II of Brassicaceae. Our analysis revealed that *M. pygmaea* is very similar to the ancestral genome PCK (Lysak et al., 2016; Schranz et al., 2006). Four chromosomes, AK1, AK4, AK7 and AK6/8, are shared between *Megadenia* and PCK. The fifth chromosome (Mp5) is similar to PCKs chromosome AK5/8/6, but differentiated by a 9.92 Mb paracentric inversion. The sixth chromosome (Mp6) was derived from ancestral chromosomes AK2 and AK3 via an end-to-end translocation (EET). EET isn one of the common mechanisms of reducing chromosome number. It usually results from two double-strand breaks (DSB) at terminal regions of two different chromosomes followed by merging the two chromosomes (Lysak, 2014). An EET event can be inferred from the retained synteny blocks corresponding to a whole ancestral chromosome without an active centromere (Mandáková & Lysak, 2018). In our study, *M. pygmaea* has a relatively simple karyotype with a single EET event such that it structurally resembles PCK but with one fewer chromosome, which most likely preceded its independent divergence and later intrageneric diversification (Artyukova et al., 2014). Further research is needed to elucidate whether an ancestral genome of *Megadenia* was directly derived from PCK or another, structurally similar, ancestral genome.

ACKNOWLEDGEMENTS

This study was supported by the National Natural Science Foundation of China (41771055, 31700323), National Key Research and Development program (2017YFC0505203), and Central European Institute of Technology (CEITEC) 2020 project (grant number LQ1601).

AUTHOR CONTRIBUTIONS

QH and JL designed the research. WY and LZ collected the materials and performed the genome sequencing and assembly. WY, LH, TL, JJ, and YY performed the genome annotation and evolution analysis. ML and TM performed the comparative chromosome painting analysis. WY, ML, JL and QH wrote the manuscript. All authors contributed to the article and approved the submitted version.

DATA AVAILABILITY STATEMENT

All raw sequence data have been deposited in the NCBI under accession number PRJNA637465. The draft genome assembly has been deposited in the Genome Warehouse in National Genomics Data Center, Beijing Institute of Genomics (China National Center for Bioinformation), Chinese Academy of Sciences, under accession number PRJCA002905.

ORCID

Yongzhi Yang https://orcid.org/0000-0001-6912-6718
Martin A. Lysak https://orcid.org/0000-0003-0318-4194
Quanjun Hu https://orcid.org/0000-0001-6922-2144

REFERENCES

Arnegard, M. E., McGee, M. D., Matthews, B., Marchinko, K. B., Conte, G. L., Kabir, S. M.,Bedford, N., Bergek, S., Frank Chan, Y., Jones, F. C., Kingsley, D. M., Peichel, C. L., & Schluter, D. (2014). Genetics of ecological divergence during speciation. *Nature*, *511*(7509), 307–311.

Artyukova, E. V., Kozyrenko, M. M., Boltenkov, E. V., & Gorovoy, P. G. (2014). One or three species in *Megadenia* (Brassicaceae): Insight from molecular studies. *Genetica*, *142*(4), 337–350. https://doi.org/10.1007/s10709-014-9778-1

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, *28*(1), 45–48. https://doi.org/10.1093/nar/28.1.45

Beilstein, M. A., Al-Shehbaz, I. A., Mathews, S., & Kellogg, E. A. (2008). Brassicaceae phylogeny inferred from phytochrome A and ndhF sequence data: Tribes and trichomes revisited. *American Journal of Botany*, *95*(10), 1307–1327.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Cheng, F., Wu, J., & Wang, X. (2014). Genome triplication drove the diversification of *Brassica* plants. *Horticulture Research*, *1*(1), 1–8. https://doi.org/10.1038/hortres.2014.24

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676. https://doi.org/10.1093/bioinformatics/bti610

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: A computational tool for the study of gene family evolution. *Bioinformatics*, *22*(10), 1269–1271. https://doi.org/10.1093/bioinformatics/btl097

Dorofeyev, V. I. (2004). System of family Cruciferae B. *Juss*.(*Brassicaceae Burnett*). *Turczaninowia*, *7*(3), 43–52.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., & Aiden, A. P. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95.

Dudchenko, O., Shamim, M. S., Batra, S., Durand, N. C., Musial, N. T., Mostofa, R., Stamenova, E. (2018). The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. Biorxiv, 254797.

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems*, *3*(1), 99–101. https://doi.org/10.1016/j.cels.2015.07.012

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems*, *3*(1), 95–98. https://doi.org/10.1016/j.cels.2016.07.002

Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, *9*(1), 18. https://doi.org/10.1186/1471-2105-9-18

Emms, D. M. (2018). STAG: Species Tree Inference from All Genes. *BioRxiv*, 267914.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1832-y

Geiser, C., Mandáková, T., Arrigo, N., Lysak, M. A., & Parisod, C. (2016). Repeated whole-genome duplication, karyotype reshuffling, and biased retention of stress-responding genes in buckler mustard. *The Plant Cell*, *28*(1), 17–27. https://doi.org/10.1105/tpc.15.00791

German, D. A., & Al-Shehbaz, I. A. (2008). Five additional tribes (Aphragmeae, Biscutelleae, Calepineae, Conringieae, and Erysimeae) in the Brassicaceae (Cruciferae). *Harvard Papers in Botany*, *13*(1), 165–170. https://doi.org/10.3100/1043-4534(2008)13[165:FATABC]2.0.CO;2

Guo, X., Liu, J., Hao, G., Zhang, L., Mao, K., Wang, X., Zhang, D., Ma, T., Hu, Q., Al-Shehbaz, I. A., & Koch, M. A. (2017). Plastome phylogeny and early diversification of Brassicaceae. *BMC Genomics*, *18*(1), 176. https://doi.org/10.1186/s12864-017-3555-3

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K. Jr, Hannick, L. I., & Town, C. D. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, *31*(19), 5654–5666. https://doi.org/10.1093/nar/gkg770

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B. O., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494. https://doi.org/10.1038/nprot.2013.084

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, *9*(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7

Hedges, S. B., Dudley, J., & Kumar, S. (2006). TimeTree: A public knowledge-base of divergence times among organisms. *Bioinformatics*, *22*(23), 2971–2972. https://doi.org/10.1093/bioinformatics/btl505

Huang, C.-H., Sun, R., Hu, Y. I., Zeng, L., Zhang, N., Cai, L., Zhang, Q., Koch, M. A., Al-Shehbaz, I., Edger, P. P., Pires, J. C., Tan, D.-Y., Zhong, Y., & Ma, H. (2016). Resolution of brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution*, *33*(2), 394–412. https://doi.org/10.1093/molbev/msv226

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, *40*(D1), D109–D114. https://doi.org/10.1093/nar/gkr988

Kang, M., Wu, H., Yang, Q., Huang, L. I., Hu, Q., Ma, T., Li, Z., & Liu, J. (2020). A chromosome-scale genome assembly of *Isatis indigotica*, an important medicinal plant used in traditional Chinese medicine. *Horticulture Research*, *7*(1), 1–10. https://doi.org/10.1038/s41438-020-0240-5

Kang, Y. J., Kim, S. K., Kim, M. Y., Lestari, P., Kim, K. H., Ha, B.-K., Jun, T. H., Hwang, W. J., Lee, T., Lee, J., Shim, S., Yoon, M. Y., Jang, Y. E., Han, K. S., Taeprayoon, P., Yoon, N. A., Somta, P., Tanya, P., Kim, K. S., … Lee, S.-H. (2014). Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nature Communications*, *5*, 5443. https://doi.org/10.1038/ncomms6443

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kiefer, M., Schmickl, R., German, D. A., Mandáková, T., Lysak, M. A., Al-Shehbaz, I. A., Franzke, A., Mummenhoff, K., Stamatakis, A., & Koch, M. A. (2014). BrassiBase: Introduction to a novel knowledge database on Brassicaceae evolution. *Plant and Cell Physiology*, *55*(1), e3. https://doi.org/10.1093/pcp/pct158

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736.

Lv, H., Fang, Z., Yang, L., Zhang, Y., & Wang, Y. (2020). An update on the arsenal: Mining resistance genes for disease management of *Brassica* crops in the genomic era. *Horticulture Research*, *7*(1), 1–18.

Lysak, M. A. (2014). Live and let die: Centromere loss during evolution of plant chromosomes. *New Phytologist*, *203*(4), 1082–1089.

Lysak, M. A., Mandáková, T., & Schranz, M. E. (2016). Comparative paleogenomics of crucifers: Ancestral genomic blocks revisited. *Current Opinion in Plant Biology*, *30*, 108–115.

Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics*, *20*(16), 2878–2879.

Mandáková, T., Guo, X., Özüdoğru, B., Mummenhoff, K., & Lysak, M. A. (2018). Hybridization-facilitated genome merger and repeated chromosome fusion after 8 million years. *Plant Journal*, *96*(4), 748–760.

Mandáková, T., & Lysak, M. A. (2008). Chromosomal phylogeny and karyotype evolution in x= 7 crucifer species (Brassicaceae). *The Plant Cell*, *20*(10), 2559–2570.

Mandáková, T., & Lysak, M. A. (2016a). Painting of *Arabidopsis* chromosomes with chromosome-specific BAC clones. *Current Protocols in Plant Biology*, *1*(2), 359–371.

Mandáková, T., & Lysak, M. A. (2016b). Chromosome preparation for cytogenetic analyses in *Arabidopsis*. *Current Protocols in Plant BiologyPlant Biology*, *1*(1), 43–51.

Mandáková, T., & Lysak, M. A. (2018). Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology*, *42*, 55–65.

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770.

Nikolov, L. A., Shushkov, P., Nevado, B., Gan, X., Al-Shehbaz, I. A., Filatov, D., Bailey, C. D., & Tsiantis, M. (2019). Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytologist*, *222*(3), 1638–1651. https://doi.org/10.1111/nph.15732

Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., & Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, *327*(5961), 92–94.

Ou, S., & Jiang, N. (2018). LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology*, *176*(2), 1410–1422. https://doi.org/10.1104/pp.17.01310

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One*, *5*(3). https://doi.org/10.1371/journal.pone.0009490

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics*, *21*, i351–i358.

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1), 1–10.

Schranz, M. E., Lysak, M. A., & Mitchell-Olds, T. (2006). The ABC's of comparative genomics in the Brassicaceae: Building blocks of crucifer genomes. *Trends in Plant Science*, *11*(11), 535–542. https://doi.org/10.1016/j.tplants.2006.09.002

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*.

Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Research*, *32*(suppl_2), W309–W312. https://doi.org/10.1093/nar/gkh379

Stebbins, G. L. (1971). *Chromosomal evolution in higher plants*. Edward Arnold Ltd.

Sun, H., Ding, J., Piednoël, M., & Schneeberger, K. (2018). findGSE: Estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics*, *34*(4), 550–557. https://doi.org/10.1093/bioinformatics/btx637

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, *25*(1), 4–10. https://doi.org/10.1002/0471250953.bi0410s25

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., Dekker, J., & Lander, E. S. (2010). Hi-C: A method to study the three-dimensional architecture of genomes. *Journal of Visualized Experiments*, *39*, 1–7. https://doi.org/10.3791/1869

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, *9*(11). https://doi.org/10.1371/journal.pone.0112963

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., Lee, T.-H., Jin, H., Marler, B., Guo, H., Kissinger, J. C., & Paterson, A. H. (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, *40*(7), e49. https://doi.org/10.1093/nar/gkr1293

Xu, Z., & Wang, H. (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, *35*(suppl_2), W265–W268. https://doi.org/10.1093/nar/gkm286

Zdobnov, E. M., & Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, *17*(9), 847–848. https://doi.org/10.1093/bioinformatics/17.9.847

Zhang, L., Cai, X. U., Wu, J., Liu, M., Grob, S., Cheng, F., Liang, J., Cai, C., Liu, Z., Liu, B. O., Wang, F., Li, S., Liu, F., Li, X., Cheng, L., Yang, W., Li, M.-H., Grossniklaus, U., Zheng, H., & Wang, X. (2018). Improved *Brassica rapa* reference genome by single-molecule sequencing and chromosome conformation capture technologies. *Horticulture Research*, *5*(1), 1–11. https://doi.org/10.1038/s41438-018-0071-9

Zhou, T. Y. (2001). Brassicaceae. *Flora of China*, *8*, 1–193.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Yang W, Zhang L, Mandáková T, et al. The chromosome-level genome sequence and karyotypic evolution of *Megadenia pygmaea* (Brassicaceae). *Mol Ecol Resour*. 2021;21:871–879. https://doi.org/10.1111/1755-0998.13291