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 Lysak[2], Jianquan Liu[1], and Quanjun Hu[1]

[1]Sichuan University
[2]Masaryk University
[3]Lanzhou University

July 16, 2020

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 pseudo-chromosomes. We annotated a total of 25,607 high-confidence protein-coding genes and corroborated the phylogenetic affinity of Megadenia with the expanded Lineage II, which contains 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 Brassicaceae Lineage II. However, an end-to-end translocation between two ancestral chromosomes reduced the chromosome number from n = 7 to n = 6, comparable to Megadenia. Our reference genome provides fundamental information for use in horticulture, plant breeding and evolutionary study of this genus.

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 structure, which can independently, or in combination, promote evolutionary divergence (Arnegard et al., 2014). The rapid diversification of Brassicaeae arose not only by polyploidy, but through karyotypic changes, providing a useful model system to study the diverse forms of karyotypic evolution (Lysak et al., 2016; Mandáková & Lysak, 2008). The Brassicaeae are a large family comprised of ca. 350 genera and nearly 4,000 species of angiosperm (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 (*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 Lineage I and II (Lysak, 2014). Comparative genomics and chromosome painting revealed that the ancestral karyotype of the Lineage I, the Ancestral Crucifer Karyotype (ACK), was comprised of 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 et al., 2018; Mandáková & Lysak,

2008), was found to be derived from the ancestral PCK (ancPCK, n = 8) through descending dysploidy, namely a reduction in chromosome number (Geiser et al., n.d.; Mandáková et al., 2018).

Megadenia is a genus of Brassicaeae 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 4000 m a.s.l. (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 indicates the early divergence of *Megadenia*from other members of the Lineage II (Guo et al., 2017). This study aimed to understand the structure and chromosome evolution of the*M. pygmaea* nuclear genome. This research established the detailed chromosome structure and performed a comparative analysis to closely related Brassicaceae to inform understanding of the PCK genome using a chromosome-level *de novo*genome assembly 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 likely mediated the reduction of chromosomes from the PCK-like genome (n = 7) to the extant *Megadenia* genome (n = 6). The new reference genome of *M. pygmaea* provides information for advancing the horticultural use of *Megadenia* and aids future investigations into evolutions and uniquely disjunct biogeography of this genus.

Materials and Methods

Plant Material

Leaves and young inflorescences were collected from wild *M. pygmaea* in the Ganzi Tibetan Autonomous Prefecture, Sichuan Province, China (**Fig. 1A**).

DNA Extraction and Genome Sequencing

DNA was extracted from leaf tissue and sequenced using Nanopore long reads sequencing technology (Senol et al., 2017). High-quality DNA was extracted using the Qiagen kit to construct a 1D library, and single-molecule sequencing of DNA was performed using the GridION X5 (Oxford Nanopore Technology). Following the manufacturer's data filtering and quality control, a total of 645,789 reads with an average read length of 21.1 kb was recovered (13.6 Gb) and, following assembly, the N50 was 29.9 kb and the longest read was 153.2 kb. Paired-end Illumina sequencing was for performed for error correction and K-mer analysis using Illumina's Genomic DNA Sample Preparation kit and the Illumina HiSeq X Ten system.

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 Jellyfish v2.29 (Marçais & Kingsford, 2011) and Illumina reads. Low-quality reads were filtered prior to *de novo* assembly, as previously described (Wu et al., 2019), and the assembly was performed with Canu v1.7 (Koren et al., 2017). We corrected the assembled contigs with two iterations of Pilon v1.23 (Walker et al., 2014). Contigs were anchored to chromosomes by Hi-C. The Hi-C library was prepared from 3 g of freshly ground young leaves, using liquid nitrogen and a mortar and pestle. The chromatin extraction, digestion, DNA ligation, purification, and fragmentation were all performed as previously described (Louwers et al., 2009). A total of 114,431,960 Hi-C Illumina reads were generated using Illumina Hiseq X Ten system. The draft assembly was scaffolded with Hi-C data using the 3D-DNA pipeline v180922 run with default parameters (Dudchenko et al., 2017). Hi-C reads were aligned to the draft assembly using the Juicer pipeline v1.6.2 (Durand, Robinson, et al., 2016; Durand, Shamim, et al., 2016). Results were polished using the Juicebox Assembly Tools - an assembly-specific module in the Juicebox visualization system v1.11.08 (Dudchenko et al., 2018). The Hi-C scaffolding resulted in six chromosome-length super scaffolds, representing a total of 95.36% of the assembled sequence.

Repeats Annotation

Repeat elements in the *M. pygmaea* genome were identified with the help of the RepeatMasker v4.0 (Tarailo-

Graovac & Chen, 2009) and RepeatModeler v4.07 (Smit & Hubley, 2011) 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 nt and the length between two LTRs set to 1,000-2,0000 nt. Results were combined using LTR_retriever v1.9 (Ou & Jiang, 2018) and the insertion time (T) was calculated for each LTR retrotransposon (T = K/2r, K: genetic distance) with a substitution rate (r) of 7×10^{-9} substitutions per site per year (Ossowski et al., 2010).

Gene Prediction and Annotation

A combination of *de novo* -, homology- and transcript-based methods was used for gene prediction. A comprehensive transcriptome database was built with the PASA pipeline v2.1.0 (Haas et al., 2003). After quality filtering with Trimmomatic v0.33 (Bolger et al., 2014), a *de novo*assembly was performed on Illumina RNA-seq reads using Trinity v2.6.6 (Haas et al., 2013). Then, genome-guided transcripts were created using (1) the genome-guided mode implemented in Trinity and (2) the HISAT-StringTie pipeline v1.3.3b (Pertea et al., 2015). Homologs were predicted by mapping protein sequences from *A. thaliana, Aethionema arabicum, Arabidopsis lyrata, B. rapa* , *Capsella rubella, Carica papaya* , *Eutrema salsugineum* and*Leavenworthia alabamica* to the *M. pygmaea* genome using tblastn (E-value [?] 1e-5), and exonerate v2.4.0 was used for gene annotation (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 assignments 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 (Zdobnov & Apweiler, 2001). Predicted gene functions and metabolic pathways were annotated using Blast2GO v2.5 (Conesa et al., 2005) and the GO (Consortium, 2004) 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.

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, C. rubella, E. salsugineum, Eutrema yunnanense, L. alabamica, Raphanus raphanistrum, Sisymbrium irio* . Protein sequences from 1,356 single-copy gene families were used for phylogenetic tree construction. Gene families were constructed using the OrthoMCL v2.0.9 (Li et al., 2003) 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 was used to generate sequence alignment for protein sequences in each gene family using the default parameters (Katoh & Standley, 2013). Conserved protein sequence alignments were extracted by Gblocks v0.91b (Castresana, 2000), and the remaining variable protein alignment regions were used to construct a phylogenetic tree with RAxML v8.2.11 (Stamatakis, 2014) using the PROTGAMMALGX model. 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 burn-in of 1,000 iterations. The calibration time of divergence was obtained from the TimeTree database (Hedges et al., 2006) (http://www.timetree.org/).

Gene Family Expansion and Contraction

The expansion or contraction of orthologous gene families was determined using CAFE v 4.2 (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, number of random = 1,000.

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 degC. Chromosome spreads were prepared from fixed young flower buds containing immature anthers as previously described (Mandakova & 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).

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 labeled 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 sulfate 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 labeled 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 analyzed 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).

Results

Genome Assembly and Annotation

We generated a total of 17.2 Gb data and 13.6 Gb clean data (**Table S1**). All Nanopore subreads were corrected using canu-correct and trimmed by canu-trim for low-quality bases. The reads [?]500 bp were used to generate an initial assembly with WTDBG. We used Pilon to polish the genome assembly twice, to finally obtain a 215.4-Mb contig-scale assembly with contig N50 of 1.81 Mb. The genome contained 447 contigs, with the longest contig being 11.13 Mb in length. We then anchored these contigs into six chromosomes with Hi-C reads by 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, Fig. 1B**).

The 215.2-Mb draft *M. pygmaea* genome represents a high-quality near-complete genome assembly. A total of 1,395/1,440 plant-specific orthologs were present, indicating an estimated completeness of 96.9% (**Table S2**). The assembly size fell only slightly below estimates from K-mer analysis and flow cytometric: 259 Mb and 219 Mb, respectively (**Figs. S1 and S2**). In total, 25,607 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 vast majority of gene models were supported by complementary DNA/expressed sequence tag evidence. In our assembly, 97.03% of the genes (24,846 of 25,607) were annotated on six chromosomes, and only 2.97% (761 of 25,607) remained on scaffolds (**Table S3**). 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 S4**). The insertions of the LTR-RTs in *M. pygmaea* occurred earlier than in *A. lyrata* (**Fig. 1C**). The *M. pygmaea* genome contains a similar number of transcription factors (TFs) (1,571) as these Brassicaceae species (**Table S5** ; http://www.transcriptionfactor.org).

Phylogeny and Whole-genome Duplication

A total of 279,614 coding sequences from *M. pygmaea* and genomes representing the two Brassicaceae Lineages (I and II) were assessed, and clustered into 28,151 gene families. Species were grouped into phylogenetic lineages according to their COG gene profiles. *M. pygmaea* shared a total of 17,919 with Lineage I species and 18,018 with Lineage II, with 292 genes unique to *M. pygmaea* (**Fig. 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) (**Fig. 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 (**Fig. 2C**). *M. pygmaea* was estimated to diverge from other Lineage II genera around 27.04 (19.11-36.60) million years ago.

Gene Expansion/Contraction and Species-specific Genes in*M. pygmaeas*

A total of 37 and 202 gene families significantly ($P < 0.05$) expanded and contracted in *M. pygmaea* , respectively, of the 758 and 2,973 that significantly differed among other Lineage II genomes (**Fig. 2C**). The significantly expanded and contracted gene families contain 1,231 and 2,476 genes, respectively. The functional annotation of these genes revealed that expanded genes were involved in hyperosmotic salinity response, regulation of defense response to fungus and stomatal movement (**Table S6**). We extracted 5,504 species-specific genes from the expanded and species-specific families in the *M. pygmaea* genome. These genes were enriched signal transduction, defense response to insect and other organisms (**Table S7**).

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. 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(M. 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 (**Figs. 3, S3 and S4**).

The complete comparative chromosomal map of *M. pygmaea* , constructed by CCP, had similarities and notable differences to the structure of ancestral Brassicaceae genomes: ACK, ancPCK and PCK (**Fig. 3**). 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 (**Fig. 3B**). Chromosome Mp2 was formed by an end-to-end translocation (EET) merging ancestral chromosomes AK2 and AK3 (**Fig. 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.

Discussion

Our study produced a high-quality genome of a shade-loving plant,*M. pygmaea* , with potential horticulture use. 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 PCK's 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. Therefore, *M. pygmaea* has a relatively simple karyotype structurally resembling 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 derived from PCK or another, structurally similar, ancestral genome. A more comprehensive sampling of extant *M. pygmaea* populations would reveal more details about the closest

relatives of this genus. This newly described, high-quality *M. pygmaea* reference genome will be a valuable resource for further horticultural research and breeding, as well as for research focused on the evolutionary trajectory and biogeography of *Megadenia* .

Availability of Supporting Data

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.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (41771055, 31700323) and Central European Institute of Technology (CEITEC) 2020 project (grant no. LQ1601).

References

Arnegard, M. E., McGee, M. D., Matthews, B., Marchinko, K. B., Conte, G. L., Kabir, S., Bedford, N., Bergek, S., Chan, Y. F., 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.

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* , *28* (1), 45–48.

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.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* , *17* (4), 540–552.

Cheng, F., Wu, J., & Wang, X. (2014). Genome triplication drove the diversification of Brassica plants. *Horticulture Research* ,*1* (1), 1–8.

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.

Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* ,*32* (suppl_1), D258–D261.

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.

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., Shamim, M. S., Machol, I., Lander, E. S., & 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., Pham, M., St Hilaire, B. G., Yao, W., & 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.

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.

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.

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.*Plant Cell* , *28* (1), 17–27.

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.

Guo, X., Liu, J., Hao, G., Zhang, L., Mao, K., Wang, X., Zhang, D., Ma, T., Hu, Q., & Al-Shehbaz, I. A. (2017). Plastome phylogeny and early diversification of Brassicaceae. *BMC Genomics* , *18* (1), 176.

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

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., & Lieber, M. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* ,*8* (8), 1494.

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.

Hedges, S. B., Dudley, J., & Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms.*Bioinformatics* , *22* (23), 2971–2972.

Huang, C. H., Sun, R., Hu, Y., 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.

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.

Kang, M., Wu, H., Yang, Q., Huang, L., 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.

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. (2014). Genome sequence of mungbean and insights into evolution within *Vigna* species.*Nature Communications* , *5* , 5443.

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.

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–e3.

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.

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* , *13* (9), 2178–2189.

Louwers, M., Splinter, E., Van Driel, R., De Laat, W., & Stam, M. (2009). Studying physical chromatin interactions in plants using Chromosome Conformation Capture (3C). *Nature Protocols* ,*4* (8), 1216.

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.

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.

Ossowski, S., Schneeberger, K., Lucas-Lledo, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D., & 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.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* ,*33* (3), 290.

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.

Senol, D., Kim, J., Ghose, S., Alkan, C., & Mutlu, O. (2017). Nanopore Sequencing Technology and Tools: Computational Analysis of the Current State, Bottlenecks and Future Directions. *Pacific Symposium on Biocomputing Poster Session* .

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

Smit, A., & Hubley, R. (2011). RepeatModeler. *Institute of Systems Biology* , *1* (5).

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* ,*30* (9), 1312–1313.

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.

Stebbins, G. L. (1971). Chromosomal evolution in higher plants.

London: Edward Arnold Ltd.

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* , *25* (1), 4–10.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., & Young, S. K. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* ,*9* (11).

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B., & Guo, H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity.*Nucleic Acids Research* , *40* (7), e49–e49.

Wu, H., Ma, T., Kang, M., Ai, F., Zhang, J., Dong, G., & Liu, J. (2019). A high-quality *Actinidia chinensis* (kiwifruit) genome.*Horticulture Research* , *6* (1), 1–9.

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.

Zdobnov, E. M., & Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro.*Bioinformatics* , *17* (9), 847–848.

Zhang, L., Cai, X., Wu, J., Liu, M., Grob, S., Cheng, F., Liang, J., Cai, C., Liu, Z., Liu, B., Wang, F., Li, S., Liu, F., Li, X., Cheng, L., Yang, W., Li, M. he, 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.

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

Figure legends

Fig. 1 A. Photo of *M. pygmaea* . **B.** The Hi-C chromatin interaction map for the six chromosomes of *M. pygmaea*genome. **C.** The evolutionary dynamics of LTR retrotransposons representing intact insertions during the last 10 million years.

Fig.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: *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 *E. salsugineum, E. yunnanense, R. raphanistrum* and*S. irio* .

Fig. 3 A. Comparative karyotype based on CCP analysis showing the position of 22 genomic blocks (A–X) on six *Megadenia*chromosomes (Mp1-Mp6). Color coding reflects the position of genomic blocks on the eight chromosomes in ACK. The *A. thaliana* BAC clones delimiting each block are shown. **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.

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
Number of genes	25 607
Average transcript length (bp)	2 628
Average CDS length (bp)	234
Average exons per gene	5.4
Average exon length (bp)	281
Average intron length (bp)	233
GC content (%)	37.1
Gap content (%)	0.2
Transposable elements (%)	42.6