Huang et al. *Horticulture Research* (2020)7:197
https://doi.org/10.1038/s41438-020-00422-w

Horticulture Research
www.nature.com/hortres

ARTICLE Open Access

A chromosome-scale reference genome of *Lobularia maritima*, an ornamental plant with high stress tolerance

Li Huang[1], Yazhen Ma[1], Jiebei Jiang[1], Ting Li[1], Wenjie Yang[1], Lei Zhang[1], Lei Wu[1], Landi Feng[1], Zhenxiang Xi[1], Xiaoting Xu[1], Jianquan Liu[1,2] and Quanjun Hu[1]

Abstract

Lobularia maritima (L.) Desv. is an ornamental plant cultivated across the world. It belongs to the family Brassicaceae and can tolerate dry, poor and contaminated habitats. Here, we present a chromosome-scale, high-quality genome assembly of *L. maritima* based on integrated approaches combining Illumina short reads and Hi–C chromosome conformation data. The genome was assembled into 12 pseudochromosomes with a 197.70 Mb length, and it includes 25,813 protein-coding genes. Approximately 41.94% of the genome consists of repetitive sequences, with abundant long terminal repeat transposable elements. Comparative genomic analysis confirmed that *L. maritima* underwent a species-specific whole-genome duplication (WGD) event ~22.99 million years ago. We identified ~1900 species-specific genes, 25 expanded gene families, and 50 positively selected genes in *L. maritima*. Functional annotations of these genes indicated that they are mainly related to stress tolerance. These results provide new insights into the stress tolerance of *L. maritima*, and this genomic resource will be valuable for further genetic improvement of this important ornamental plant.

Introduction

Whole-genome duplication (WGD), or polyploidy, has had a strong influence on the evolution of the tree of life, and it seems to have occurred in the evolutionary history of most plant species[1,2], especially in angiosperms[3]. WGDs have been found in most angiosperm families with abundant species, including Brassicaceae, Poaceae, Asteraceae, Solanaceae, Fabaceae and Orchidaceae[4–11]. Previous studies suggested that WGDs can strengthen the adaptation of plants to environmental challenges[12] because of genomic reorganization and novelties. Through subfunctionalization or reciprocal loss of duplicated genes in differentiated populations of an ancestral species, WGDs can also promote reproductive isolation and thus facilitate speciation[13]. Brassicaceae (also known as Cruciferae), a monophyletic group distributed worldwide, has been highly diversified by complicated WGD events and subsequent evolution, with ~350 genera and 4000 species[14,15]. It contains many important crops (e.g., cabbage, rapeseed and mustard) that have been domesticated for food, biofuels, and ornamentals[16]. The well-known model organism *Arabidopsis thaliana*, which is of paramount importance in studies of the development, gene expression and genome evolution of flowering plants, is also a member of this family[17,18]. Analyses of the *A. thaliana* genome have provided clear evidence that three ancient WGD events (γ, β and α), occurred in its evolutionary history. The oldest WGD event, the At-γ event, was related to the diversification of eudicots and perhaps all angiosperms[19–21]. The At-β event postdated the Brassicaceae–Caricaceae divergence ~70 million years ago

Correspondence: Quanjun Hu (huquanjun@gmail.com)
[1]Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, 610065 Chengdu, China
[2]State Key Laboratory of Grassland Agro-Ecosystem, Institute of Innovation Ecology, Lanzhou University, Lanzhou, China
These authors contributed equally: Li Huang, Yazhen Ma

© The Author(s) 2020
Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Downloaded from academic.oup.com/hr/article/doi/10.1038/s41438-020-00422-w/6445704 by guest on 03 September 2026

(Mya)[22,23]. However, the At-α event was specific to the Brassicaceae family[19], occurring ~40 Mya[24]. In addition, independent WGDs more recent than the Neogene may have promoted the colonization of harsh environments by Brassicaceae taxa by increasing their stress tolerance and conferring high adaptability[25,26]. However, detailed investigation of WGDs in numerous genera present in arid habitats is still badly needed[27,28].

Lobularia maritima (L.) Desv., commonly known as sweet alyssum, is a perennial and diploid (2n = 24) herbaceous plant of the family Brassicaceae. This ornamental plant naturally occurs in the western Mediterranean region and has been widely cultivated since its domestication[29,30]. Its flowers range in color from pale violet to deep purple[31]. In addition to tolerating dry and poor habitats, *L. maritima* is recognized as a nickel hyperaccumulator that can remove heavy metals from contaminated soils[32]. As a facultative halophyte closely related to *Arabidopsis thaliana*, *L. maritima* seems to be an ideal model for revealing the molecular mechanisms underlying plant tolerance to drought and salt stress[33]. However, studies of *L. maritima* have focused mainly on its cultivation, management and rapid propagation in vitro[29].

In this study, we report a chromosome-scale assembly of the *L. maritima* genome anchored on 12 pseudochromosomes. We further identified a recent *L. maritima*-specific WGD event that occurred after the Brassicaceae-specific At-α event using comparative and evolutionary analyses. We also revealed numerous genomic changes by which *L. maritima* has adapted to harsh habitats.

Results

Genome sequencing and assembly

Samples for genome sequencing were obtained from an *L. maritima* seedling with purple flowers (Fig. 1a). We obtained 59.77 Gb of clean reads with various insert sizes and 22.31 Gb of Hi–C clean reads (~112.89-fold coverage) after Illumina sequencing and quality control (Supplementary Table 1). Two methods were employed to estimate the genome size of *L. maritima*. First, we determined the *L. maritima* genome size to be 225 Mb using flow cytometry with *A. thaliana* as the external control (Supplementary Fig. 1). Second, we used *k*-mer-based statistics[34], and the genome size was calculated to be 264 Mb (Supplementary Fig. 2).

Based on the clean reads, a de novo genome was assembled with a 197.70 Mb length. We further anchored this genome on 12 pseudochromosomes (Fig. 1b and Table 1 and Supplementary Fig. 3). We then evaluated the completeness of this genome using BUSCO v4.1.2[35] and found that 99% of the single-copy orthologs were intact (Supplementary Table 2), suggesting the high quality of the assembled genome.

Genome annotation

To predict protein-coding sequences, we combined de novo and homology- and transcriptome-based methods. We predicted 25,813 complete protein-coding genes. Gene length and the number of exons of these protein-coding genes were 2431 base pairs (bp) and 5.46 exons, respectively, on average (Table 1). In our assembly, 97.99% (25,295 of 25,813) of the genes were annotated on 12 pseudochromosomes, and only 2.01% (518 of 25,813) were located on scaffolds. The Circos v0.69 (http://circos.ca) was used to visualize the collinearity blocks between *L. maritima* and *Capsella rubella*, gene density, *Copia* density, and *Gypsy* density on individual chromosomes (Fig. 1c). Among the 25,813 predicted genes, 81.30% and 95.71% had homologs in the Swiss-Prot[36] and TrEMBL[36] databases, respectively. Additionally, we annotated 95.15%, 80.65%, and 36.55% of the genes using the InterPro[37], Gene Ontology (GO)[38] and Kyoto Encyclopedia of Genes and Genomes (KEGG)[39] databases, respectively (Supplementary Table 3). In addition, 41.94% (83 Mb) of the assembled *L. maritima* genome comprised repetitive sequences (Supplementary Table 4). Of these repetitive sequences, long terminal repeat (LTR) retrotransposons were the most frequent, spanning 14.24% of the assembled genome with 13.23% intact LTR retrotransposons. The other common repetitive sequences were DNA transposons (10.06%), Tandem Repeats (8.56%) and LINEs (5.65%) (Supplementary Tables 4 and 5). To analyze the evolutionary dynamics of these LTRs, we estimated their insertion dates in four related species (*A. thaliana*, *Arabidopsis lyrata*, *C. rubella* and *L. maritima*). The recent insertions in *A. lyrata* may have contributed to its relatively large genome size (207 Mb). Similarly, *L. maritima* had more recent insertions than *A. thaliana* and *C. rubella* (Fig. 1d and Supplementary Table 5). Diverse genetic changes can be caused by transposable elements (including LTR retrotransposons), which might have promoted lineage-specific diversification and adaptation[40]. This may partly contribute to the tolerance of *L. maritima* to arid habitats. However, the *L. maritima* genome contained a similar number of transcription factors (1799) as the other closely related Brassicaceae species (Supplementary Table 6, all transcription factor data for other species were downloaded from http://www.transcriptionfactor.org).

Comparative genomic analyses and WGD analyses

Using the ColinearScan v1.0.1[41] program and MCScanX v1[42] package, the protein sequences of *L. maritima* were compared to those of the diploid *C. rubella*, which has not been affected by a recent WGD event, to identify the collinear blocks in the genomes. The whole-genome alignments showed high collinearity and conservation, and several collinear regions almost completely spanned

Downloaded from academic.oup.com/hr/article/doi/10.1038/s41438-020-00422-w/6445704 by guest on 03 September 2026

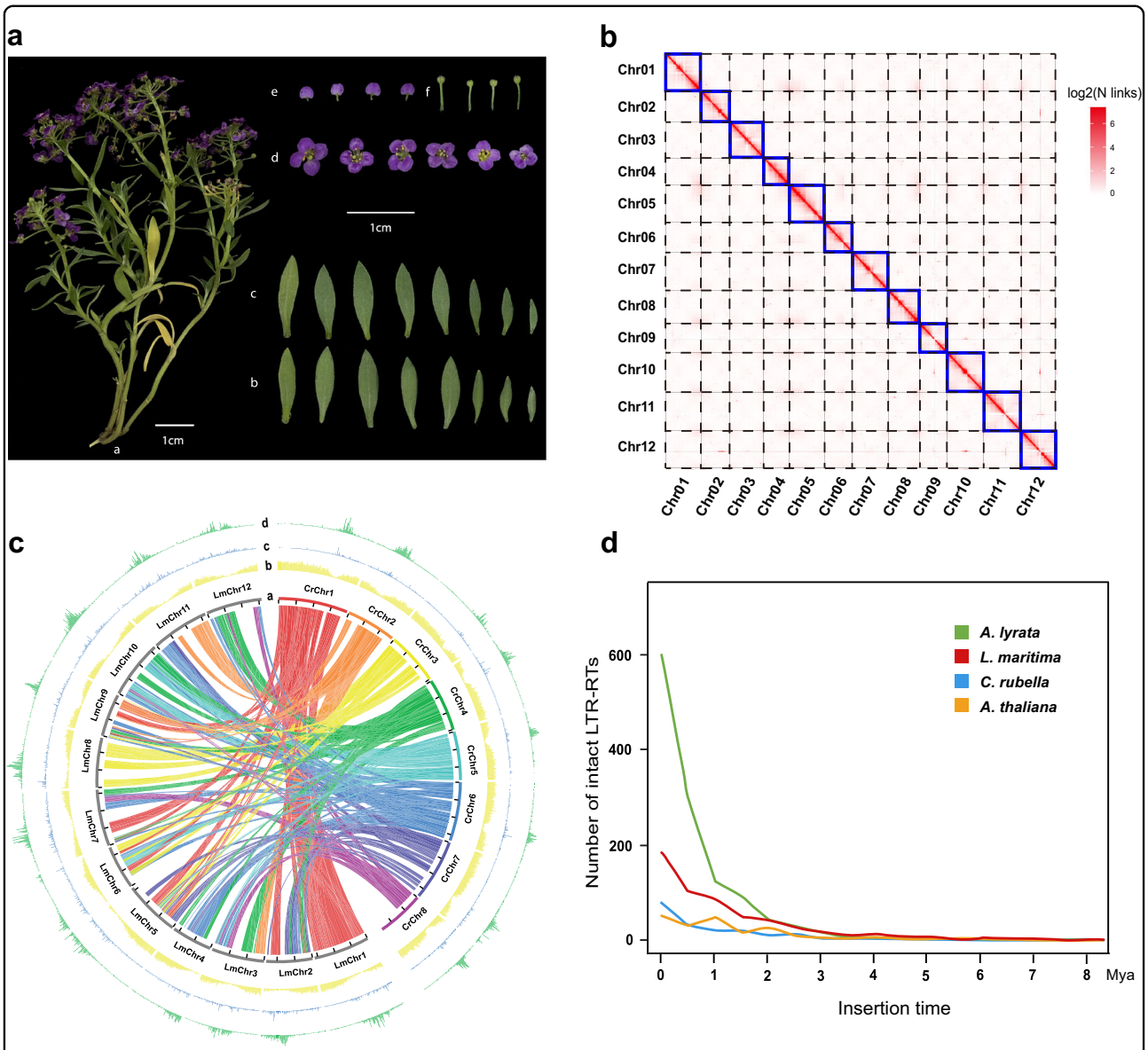

Fig. 1 Morphological and genomic characteristics of *L. maritima*. **a** Morphological characteristics of *L. maritima*: schematic representations of (a) aerial parts; (b) abaxial surfaces of leaves; (c) adaxial surfaces of leaves; (d) flowers; (e) petals; and (f) a receptacle and pedicel. **b** Hi–C chromatin interaction map for the 12 pseudochromosomes of the *L. maritima* genome. **c** Genome comparison between *L. maritima* and *C. rubella* at the chromosome level: (a) syntenic relationships between the *L. maritima* and *C. rubella* genomes; (b) gene density (window size = 100 kb, nonoverlapping); (c) density distribution of *Copia* elements (window size = 100 kb, nonoverlapping); and (d) density distribution of *Gypsy* elements (window size = 100 kb, nonoverlapping). **d** Estimated insertion times of intact LTR retrotransposons

chromosomes of the two species (Fig. 2). It is worth noting that each chromosome or chromosomal region in *C. rubella* was represented on multiple independent chromosomes in the *L. maritima* genome after the Brassicaceae-specific At-α WGD event, suggesting that the *L. maritima* genome experienced a specific WGD event.

Furthermore, we determined the karyotype of *L. maritima* using previously reported methods[28,43] (Supplementary Fig. 4) and recovered two sets of conserved genomic blocks[44,45]. However, the patterns of genomic blocks suggested that *L. maritima* experienced many postpolyploid diploidization events and a reduction in chromosome number. We also analyzed the gene retention rates of the two subgenomes in each genomic block with the *C. rubella* genome as the reference. The results showed that the two subgenomes retained similar numbers of genes (Supplementary Table 7). We also

Downloaded from academic.oup.com/hr/article/doi/10.1038/s41438-020-00422-w/6445704 by guest on 03 September 2026

assessed the absence or presence of genome dominance by examining the expression levels of each pair of duplicated genes with high confidence. Based on RNA-seq data from flower, leaf, and stem tissues, we failed to find any evidence of biased expression in each genomic block between the two subgenomes (Supplementary Table 8).

Table 1 Assembly and annotation statistics of the *L. maritima* genome

Genome feature	Value
Estimated genome size (Mb)	264
Total scaffold number	27,734
Total length (bp)	197,688,650
Total length of chromosomes (bp)	174,586,151
Longest scaffold length (bp)	16,503,592
Scaffold L50	7
Scaffold N50 length (bp)	14,943,599
GC content (%)	36.02
Repeat content (%)	41.94
Number of predicted genes	25,813
Average coding sequence length (bp)	241
Average gene length (bp)	2431
Number of exons	140,984
Average number of exons per gene	5.46

These results are largely consistent with the patterns of autopolyploids, which usually show a few instances of biased gene retention and no genome dominance.

Recent WGD event in *L. maritima*

To identify possible WGD events, we calculated the Ks values between the collinear genes. The *L. maritima* collinear blocks produced two visible peaks, at 0.583 and 1.287 (Fig. 3a), representing two different WGD events. We then estimated the occurrence times of each WGD event based on the Ks values. However, dating ancestral events in plants can be influenced by divergent evolutionary rates[46]. Thus, by aligning the *L. maritima* peak with the corresponding location in the *C. rubella* Ks distribution, as in a previous report[46], we performed evolutionary rate correction (Fig. 3b). After correction, the peaks of Ks for the two WGD events were 0.378 and 0.855, corresponding to 22.99 and 52.01 Mya, respectively (Fig. 3b). The results indicated an ancient WGD event shared with *C. rubella* and a recent species-specific WGD event in *L. maritima*. In addition, Ks estimation indicated that *C. rubella* and *L. maritima* diverged approximately 21.53 Mya (Fig. 3b). These findings were consistent with those of the synteny and collinearity analyses of *L. maritima* and *C. rubella* and suggested that *L. maritima*

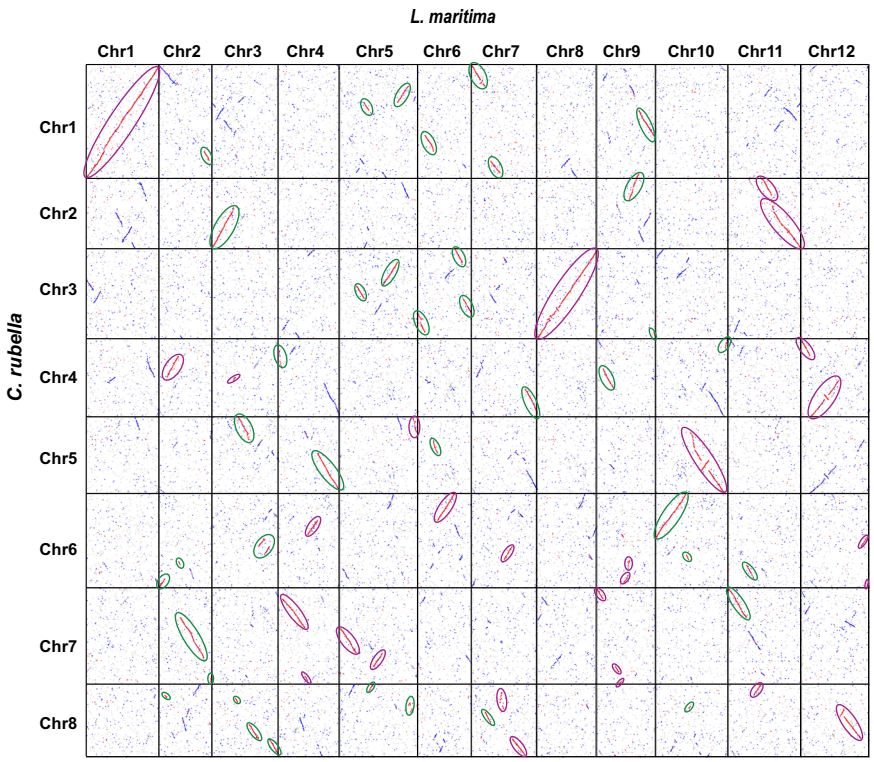

Fig. 2 Dotplot comparing the *L. maritima* and *C. rubella* genomes. Collinear regions in the *L. maritima* genome. Regions from putative subgenomes are circled in purple and green, respectively

Downloaded from academic.oup.com/hr/article/doi/10.1038/s41438-020-00422-w/6445704 by guest on 03 September 2026

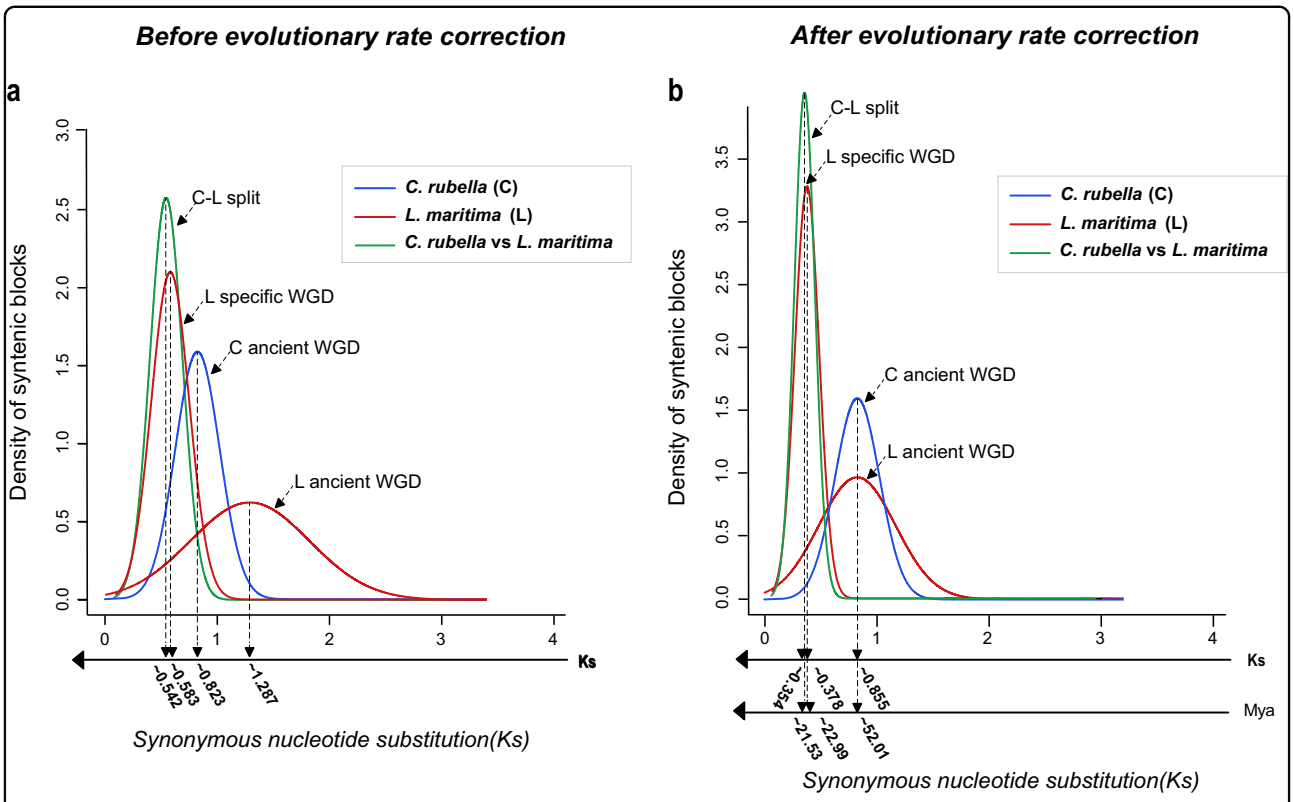

Fig. 3 Evolutionary rate correction. a Distribution of uncorrected Ks values in syntenic blocks; **b** distribution of corrected Ks values in syntenic blocks and age estimates for the events

experienced a species-specific WGD event after sharing a WGD event with other Brassicaceae species.

Phylogeny and divergence

We obtained the genome sequences of representative Brassicaceae species to clarify the genome evolution and divergence of *L. maritima*. Gene family clusters were defined based on the *L. maritima* protein-coding genes and the annotated gene sets of 10 published genomes (Supplementary Table 9) using OrthoFinder v2.3.12[47]. A total of 25,316 orthogroups were determined across the 11 species. Among these orthogroups, 1,986 were putative single-copy gene families, and 24,705 genes from *L. maritima* could be clustered into 16,821 orthogroups. In addition, we identified 1878 *L. maritima*-specific genes in these gene families. Functional annotations of these genes indicated that they were distinctly enriched in the GO terms "positive regulation of response to salt stress", "abscisic acid-activated signaling pathway", "response to freezing", "response to stimulus", and "response to biotic stimulus", indicating that the genes retained after the WGD event may be relevant in the adaptation of *L. maritima* to multiple environmental stress factors (Fig. 4a and Supplementary Tables 10 and 11). For example, a homolog of these *L. maritima*-specific genes, *ABI4*, acts as both an activator and a repressor of gene expression and plays a critical role in phytohormone signaling pathways in plant development and biotic/abiotic stress responses[48]. Another homolog, *ABI1*, serves as a key repressor of the abscisic acid (ABA) signaling pathway and regulates diverse ABA responses to abiotic stress[49,50]. The species-specific calcium-dependent protein kinase (*CDPK*) genes recovered here (Supplementary Table 11) were also demonstrated to be involved in numerous aspects of plant growth and development, from sensing biotic and abiotic stress to mediating hormone-related development[51].

To verify the phylogenetic position of *L. maritima*, we used the concatenated protein sequence alignment of the 1986 single-copy gene families in the 11-species phylogenetic analyses. The results confirmed that *L. maritima* belonged to Lineage II[11,52] (Fig. 4b), consistent with its position in the chloroplast genome phylogeny reported previously[53]. In our analyses performed using MCMCtree[54], *L. maritima* was estimated to have diverged from the other closely related species ~22.63 (18.74, 26.61) Mya (Fig. 4b).

Expansion and contraction of gene families in *L. maritima*

Gene families with significantly expanded or contracted copy numbers are usually related to the adaptive

Downloaded from academic.oup.com/hr/article/doi/10.1038/s41438-020-00422-w/6445704 by guest on 03 September 2026

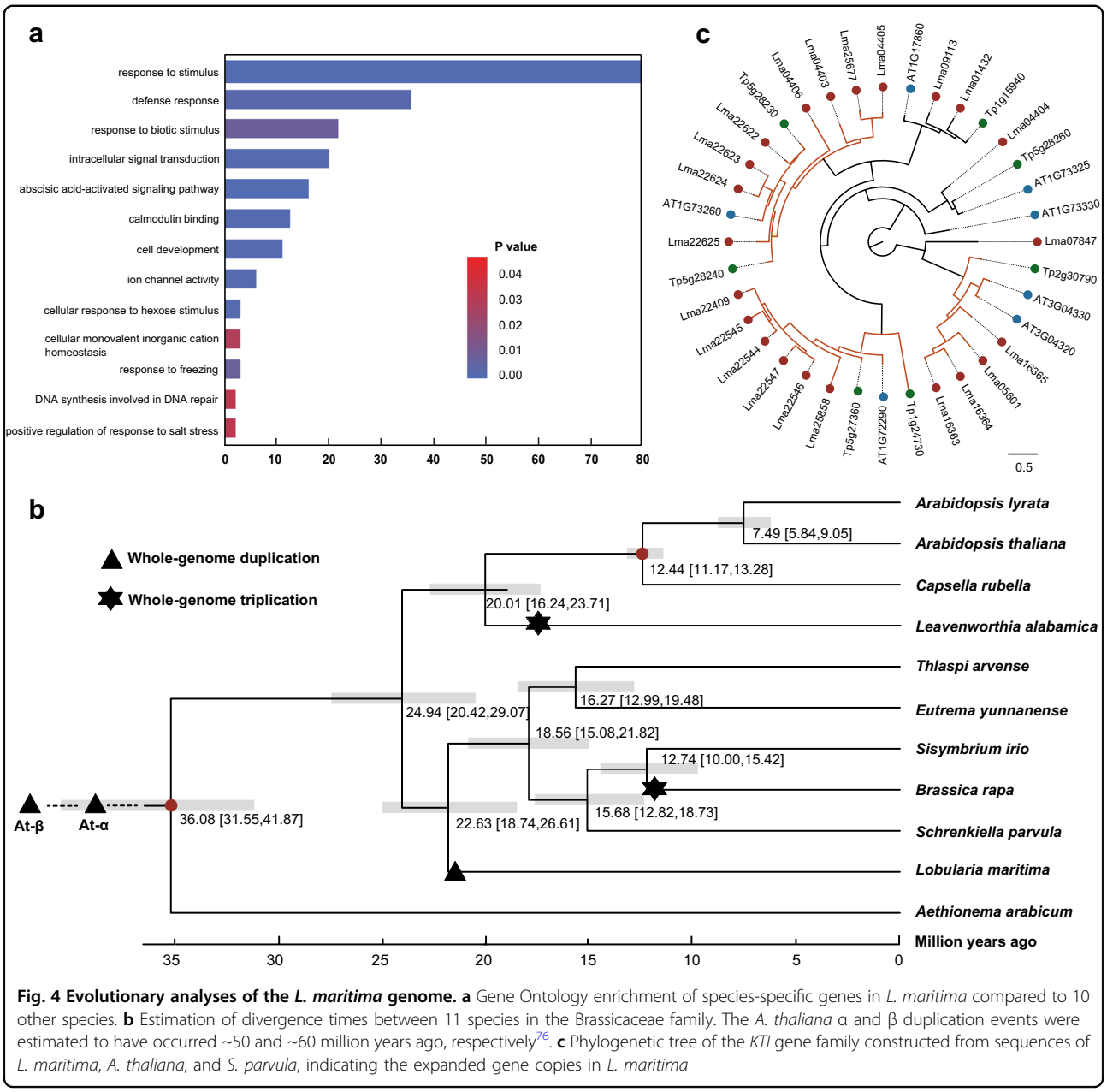

Fig. 4 Evolutionary analyses of the *L. maritima* genome. a Gene Ontology enrichment of species-specific genes in *L. maritima* compared to 10 other species. **b** Estimation of divergence times between 11 species in the Brassicaceae family. The *A. thaliana* α and β duplication events were estimated to have occurred ~50 and ~60 million years ago, respectively[76]. **c** Phylogenetic tree of the *KTI* gene family constructed from sequences of *L. maritima*, *A. thaliana*, and *S. parvula*, indicating the expanded gene copies in *L. maritima*

divergence of one species from closely related species[55,56]. We compared the genomes of *L. maritima* and 10 other species, with *Aethionema arabicum* as the outgroup (Fig. 4b), to explore the expansion and contraction of the gene families in *L. maritima*. Twenty-five gene families, comprising 319 genes, were significantly expanded in *L. maritima* ($P < 0.05$). Functional annotation of these genes indicated that they were mainly enriched in "response to molecule of bacterial origin", "response to insect", "response to molecule of fungal origin", "response to wounding" and "response to salt stress" (Supplementary Tables 12 and 13). For example, one of the expanded gene families, the *KTI* gene family, comprised versatile protease inhibitors related to defense against insect attack (Fig. 4c)[57]. In addition, the *HIPP* gene family, involved in stress responses[58], was also greatly expanded in *L. maritima*.

Positively selected genes in *L. maritima*

Genes with signs of positive selection are usually regarded to be involved in the adaptive divergence of one species from closely related species[59]. We conducted positive selection analysis by using *L. maritima* as the foreground branch and five related Brassicaceae species (*Eutrema yunnanense*, *C. rubella*, *A. arabicum*, *A. lyrata*, and *Schrenkiella parvula*) as the background branches. We identified 10,581 single-copy orthologous gene

Downloaded from academic.oup.com/hr/article/doi/10.1038/s41438-020-00422-w/6445704 by guest on 03 September 2026

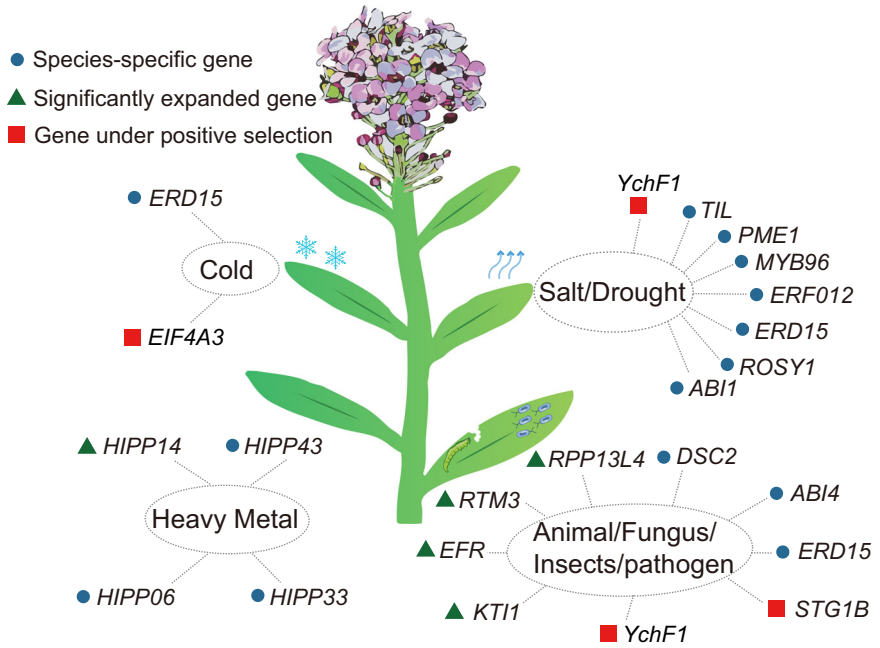

Fig. 5 Genes in the *L. maritima* genome associated with various biotic and abiotic stimuli. Gene IDs for the gene names are listed below. *ABI1*: Lma13276; *ABI4*: Lma14310, Lma25376; *EFR*: Lma05693, Lma14720, Lma21740; *EIF4A3*: Lma15462; *ERD15*: Lma14373, Lma26266; *ERF012*: Lma17999; *HIPP06*: Lma19275; *HIPP14*: Lma12975; *HIPP33*: Lma25212; *HIPP43*: Lma16268; *KTI1*: Lma22625, Lma04403, Lma25677, Lma22622, Lma22621, Lma04406, Lma22624, Lma22623, Lma25676, Lma04405; *MYB96*: Lma13875; *STG1B*: Lma21436; *PME1*: Lma13269; *ROSY1*:Lma11854; *RPP13L4*: Lma09152, Lma03536, Lma21662, Lma03543, Lma09149, Lma21664, Lma09145; *RTM3*: Lma20357, Lma22531, Lma20352, Lma10871, Lma08006, Lma20353; *TIL*: Lma24858, Lma24859; *DSC2*: Lma14716 Lma17306; *YchF1*: Lma17801; Abbreviations: ABI ABA insensitive; EFR EF-TU receptor; EIF eukaryotic initiation factor; ERD early responsive to dehydration stress; ERF ethylene-responsive factor; HIPP heavy metal-associated isoprenylated plant protein; KTI Kunitz trypsin inhibitor; MYB myeloblastosis oncogene; PME pectin methyl esterase; ROSY interactor of synaptotagmin; RPP resistance to *P. pachyrhizi*; RTM restricted tobacco etch virus movement; TIL temperature-induced lipocalin; DSC2, desmocollin-2

families. To identify the genes that evolved in response to positive selection, we adopted the branch-site model in the PAML v4.9 package[54]. After false discovery rate (FDR) correction, we identified 50 genes that were possibly under positive selection. The functions of the significantly positively selected genes (PSGs) indicated that they were associated with stress tolerance and the survival of plants (Fig. 5 and Supplementary Table 14). For example, one of the genes was *SGT1B*, which was found to be involved in innate immunity and resistance in plants mediated by multiple R genes[60–63]. Another of the genes was *YchF1*, which is involved in salinity stress tolerance and disease resistance against bacterial pathogens[64]. Another of the genes, *EIF4A3*, is an important factor for abiotic stress adaptation, which can regulate plant resistance to abiotic stress partially by regulating the expression of acetoacetyl-CoA thiolase 2[65].

Discussion

L. maritima is an important ornamental plant in horticulture because of its colorful flowers and stress tolerance. In this study, by combining Illumina and Hi–C data, a chromosome-level high-quality *L. maritima* genome was assembled. The *L. maritima* genome was ~197.70 Mb in size, and 88.31% (174.59 Mb) of the sequences were assigned to 12 pseudochromosomes. We annotated 25,813 genes and found substantially more repetitive elements (especially intact LTR retrotransposons) in the *L. maritima* genome than in the genomes of other Brassicaceae species. In addition, most intact LTR retrotransposons expanded rapidly in the recent past. Such proliferation of LTR retrotransposons may have partly resulted in the increased genome size of *L. maritima*. Phylogenetic reconstructions showed that *L. maritima* diverged early as an independent branch of Brassicaceae Lineage II.

In the histories of many diverse eukaryotes, including *Danio rerio*[66], *Saccharomyces cerevisiae*[67], and *A. thaliana*[68–71], WGDs have been discovered. Through large-scale phylogenomic analyses, ancient WGDs were found to occur in the common ancestors of both seed plants and angiosperms[4,9,71,72]. WGDs have played an essential role in angiosperm diversification and environmental adaptation[9]. Polyploids can tolerate high environmental stress, with present-day polyploids often appearing to occur at high frequencies in disturbed and harsh environments[73–75].

Downloaded from academic.oup.com/hr/article/doi/10.1038/s41438-020-00422-w/6445704 by guest on 03 September 2026

Under environmental stresses, polyploids may have been more successful because their changing environments created many opportunities to make use of the evolutionary benefits of WGDs[76]. The comparison of *L. maritima* and the diploid *C. rubella* indicated a recent WGD event that was specific to *L. maritima*, followed by extensive chromosomal rearrangements. Furthermore, we evaluated whether biased gene retention occurred after the WGD event. Two subgenomes retained a similar number of genes. However, neither subgenome showed genome dominance. This indicates that *L. maritima* might have undergone an autopolyploidization event. Analysis of the Ks values between the collinear genes suggested that the recent *L. maritima*-specific WGD event occurred ~22.99 Mya. The comparison of between-species Ks distributions indicated that the *L. maritima*-*C. rubella* divergence occurred ~21.53 Mya. Thus, this divergence and the aforementioned *L. maritima*-specific WGD event occurred at almost the same time. *L. maritima* and *C. rubella* belong to two major lineages, and it is highly likely that the divergence of the two major lineages and genus diversification of each lineage in Brassicaceae occurred radiatively at the same time. This rapid radiation was accompanied by polyploidy in a few of the genera. This is also consistent with the previous suggestion that further WGDs might have occurred in Brassicaceae since the Neogene, with radiative diversification, which further helped members of this family colonize arid habitats by increasing their stress tolerance[26]. As a result of the WGD event, species-specific genes and expanded gene families become further involved in responses to environmental stresses, for example, drought and pathogen attack, which might have facilitated the adaptation of *L.* maritima to harsh environments. In addition, the positively selected genes in *L. maritima* may have increased defense against fungal and bacterial attack. Thus, the species-specific WGD event may have promoted the adaptation of *L. maritima* to harsh environments, which is consistent with previous findings for numerous plants[76,77]. These genomic traits may also explain why *L. maritima* is a nickel hyperaccumulator[32] and a halophyte with a high tolerance to salt stress[33]. Overall, whole-genome sequencing of *L. maritima* could elucidate the stress tolerance of this ornamental plant and be useful in future breeding programs.

Materials and methods

Materials and DNA/RNA extraction

The *L. maritima* seedling was cultivated in Jinjiang District, Chengdu City, Sichuan Province, China (N 30°34′ 21.86″, E 104°09′45.47″). We harvested fresh and healthy roots, stems, leaves and flowers and immediately froze them in liquid nitrogen. Before DNA/RNA extraction, we stored these tissues in a −80 °C freezer in the laboratory. To extract high-quality genomic DNA, the cetyl trimethylammonium bromide (CTAB)[78] method was used. Additionally, we extracted total RNA from the flower, stem and leaf tissues using Qiagen RNeasy Plant Mini Kits.

Library construction and sequencing

We randomly fragmented the purified genomic DNA using a focused ultrasonicator and obtained fragments of desired lengths by electrophoresing the DNA fragments in 0.8% General Purpose Agarose E-Gel. Then, we created Illumina libraries with large (2-, 5-, 10- and 20-kb) and small (350- and 500-bp) inserts using the purified DNA fragments. Based on the PE-150 protocol, the libraries were finally sequenced on an Illumina HiSeq 2000 platform. RNA libraries were constructed with a TruSeq RNA Library Preparation Kit v2 and sequenced on the same platform.

A Hi–C library was constructed using five main steps. First, we fixed the sample with formaldehyde and crosslinked DNA-DNA interactions that are bridged by proteins. Second, the crosslinked DNA was treated with the restriction endonuclease Hind III to produce sticky ends. Third, terminal DNA repair was used to introduce biotin-labeled bases in order to facilitate subsequent DNA purification and capture. Next, we ensured the location of the interacting DNA through cyclization of the end-repaired DNA and DNA fragments. Finally, we extracted and purified the DNA sample and then used Covaris S2 to shear the DNA sample. After A-tailing, pulldown, and adapter ligation, the DNA library was sequenced on an Illumina platform using the PE-150 protocol. We used HiCPro v2.8.1[79] to remove duplicates and then assessed quality. After trimming low-quality reads and removing adapters, more than 22.31 Gb (~112.89-fold coverage) of clean data was generated. Then, all clean data were submitted to the 3D-DNA v180419 pipeline[80].

Genome assembly

Approximately 79.49 Gb of raw reads was generated by sequencing all six DNA libraries. These raw reads were filtered following a previous study[81]. We first used Trimmomatic v0.33[82] to perform quality filtering of short reads. We then used the BFC error corrector[83] followed by FastUniq v1.1[84] to delete duplicates in the mate pair data. The resultant reads produced approximately 59.77 Gb of clean data (Supplementary Table 1).

We used Platanus v1.2.4[85] software to perform *de novo* assembly of the *L. maritima* genome. Thereafter, using the 3D-DNA v180419 pipeline[80], the draft assembly was scaffolded with the Hi–C clean reads. Using the Juicer v1.6.2 pipeline[86], we aligned the Hi–C clean reads to the draft assembly genome. We then used Juicebox Assembly Tools[87] to polish the results from the 3D-DNA v180419 pipeline. The Hi–C scaffolding was anchored on 12 pseudochromosomes. In total, 88.31% of the assembled

Downloaded from academic.oup.com/hr/article/doi/10.1038/s41438-020-00422-w/6445704 by guest on 03 September 2026

sequences were related to the pseudochromosomes. In addition, we assessed the quality of the assembled genome using the BUSCO v4.1.2[35] pipeline (database: embryophyta odb10, 2020-09-02, containing 1,614 BUSCO genes).

Repeat element annotation

Repeat elements were identified with the RepeatMasker v4.0.7[88] and RepeatModeler v1.0.11[89] programs using the assembled *L. maritima* genome as the input. We also identified intact LTR retrotransposons by searching the *L. maritima* genome using LTRharvest v1.5.10[90] and LTR_Finder v1.06[91]. We further combined these results using LTR_retriever v1.9[92]. We also estimated insertion time according to a substitution rate of 7×10^{-9}/site/year.

Gene prediction and annotation

To predict genes in the *L. maritima* genome, we first assembled transcripts using the *de novo* and genome-guided modes in Trinity v2.6.6[93]. Then, these transcripts were used to create transcript-based predictions with the PASA v2.1.0 pipeline[94]. We also carried out homolog predictions. In such predictions, the protein sequences of *A. thaliana*, *A. arabicum*, *A. lyrata*, *Eutrema yunnanense*, *Brassica rapa*, *Sisymbrium irio*, *C. rubella*, *Tarenaya hassleriana*, *Leavenworthia alabamica* and *Carica papaya* were mapped to the *L. maritima* genome using Exonerate v2.2.0 (https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate). GlimmerHMM v3.0.4[95] and Augustus v3.2.2[96] were trained with genes from the PASA results and used for *de novo* gene prediction. We merged the gene models from the three sources using EVidenceModeler v1.1.1[97]. To annotate the functions of all predicted genes, we aligned the protein sequences of *L. maritima* to Swiss-Prot and TrEMBL[36] using blastp and generated functional assignments based on the best hit. Protein domains were determined by searching against the InterPro[37] database. In addition, Blast2GO v2.5[98] was used to identify the Gene Ontology[38] annotations and KEGG[39] pathways using the KAAS server (https://www.genome.jp/kegg/kaas).

Synteny and WGD

To construct syntenic blocks between *L. maritima* and *C. rubella*, all protein sequences of *L. maritima* were compared to protein sequences of *C. rubella*. The gene pairs with an *e*-value ≤ 1e-5 were further analyzed. We applied the ColinearScan v1.0.1[41] program, which can effectively evaluate genomic blocks of collinear genes, and the MCScanX v1 package[42] to find the syntenic blocks between the *C. rubella* and *L. maritima* genomes. Thereafter, we used these collinear gene pairs to construct a dotplot. Next, we used the script "add_ka_and_ks_to_collinearity.pl" in MCScanX to calculate the Ks values of the collinear orthologous gene pairs. We converted the Ks values to divergence times (T) based on $T = Ks/2r$, where r is the neutral substitution rate (8.22×10^{-9}). Finally, we performed evolutionary rate correction because of the inconsistent evolutionary rates among species. The evolutionary rate correction method was as reported by Wang et al.[46]. Briefly, under the assumption that the *C. rubella* peak appears at k_C and the *L. maritima* peak appears at k_L, we can use the equation $r = (k_L - k_C)/k_C$ to describe the relative evolutionary rate of *L. maritima*. Then, rate correction was performed to discover the corrected rate $k_{L\ \mathrm{correction}}$ of *L. maritima* relative to k_C: (1) For the Ks between duplicates in *L. maritima*, we defined the correction coefficient W_L as $k_{\mathrm{L\ correction}}/k_L = k_C/k_L = W_L$; thus, we obtained $k_{\mathrm{L\ correction}} = k_C/k_L \times k_L = 1/(1 + \mathrm{r}) \times k_L$ and $W_L = 1/(1 + \mathrm{r})$. (2) For the Ks between homologous genes from *C. rubella* and *L. maritima*, if the peak was located at $k_{L\text{-}C}$, supposing the correction coefficient W_L in *L. maritima*, we then calculated a corrected evolutionary rate $k_{\mathrm{L\text{-}C\text{-}correction}} = W_L \times \mathrm{k}_{L\text{-}C}$.

Phylogeny and divergence

The genomes of *L. maritima* and 10 other species (*A. arabicum*, *B. rapa*, *L. alabamica*, *E. yunnanense*, *S. irio*, *A. thaliana*, *A. lyrata*, *C. rubella*, *S. parvula* and *Thlaspi arvense*) were selected to generate clusters of gene families. We retained only the longest protein sequence. We removed redundant sequences based on alternative splicing variations. Using OrthoFinder v2.3.12[47], we obtained orthologous gene families. Protein sequences from 1986 single-copy gene families were used to construct a phylogenetic tree. MAFFT v7.313[99] software was used for sequence alignment of each single-copy gene family with default settings. A phylogenetic tree was built using RAxML v8.0.0[100] under the PROTGAMMALGX model, and divergence times were calculated using the MCMCTree program of the PAML v4.9 package[54]. The calibration information for MCMCTree was extracted based on the TimeTree database[101] (http://www.time.org/).

Gene family expansion and contraction

Based on the dated phylogeny, we determined the expansions and contractions of orthologous gene families in the 11 Brassicaceae species (*A. arabicum*, *B. rapa*, *L. alabamica*, *E. yunnanense*, *S. irio*, *T. arvense*, *C. rubella*, *A. thaliana*, *A. lyrata*, *S. parvula*, and *L. maritima*) by using the CAFÉ v4.2[102] program. Genes in significantly expanded families were then used for Gene Ontology enrichment analysis.

Genes under positive selection

We selected six genomes, i.e., those of *A. arabicum*, *A. lyrata*, *C. rubella*, *E. yunnanense*, *S. parvula* and *L. maritima*, to identify orthologs for analyzing positive selection. First, Proteinortho v6.0.21[103] was used to detect

Downloaded from academic.oup.com/hr/article/doi/10.1038/s41438-020-00422-w/6445704 by guest on 03 September 2026

orthologs among the six genomes. Next, we used the PosiGene v0.1[104] pipeline for genome-wide detection of the genes with positive selection and specified the *L. maritima* clade as the foreground branch. Finally, PSGs were identified based on an FDR-corrected P value < 0.05.

Acknowledgements
This work was supported by the National Key Research and Development Program of China (2017YFC0505203), National Natural Science Foundation of China (31590821), Fundamental Research Funds for the Central Universities (2018CDDY-S02-SCU and SCU2019D013), National High-Level Talents Special Support Plan (10 Thousand of People Plan), and 985 and 211 Projects of Sichuan University.

Author contributions
Q.H., J.L., and Z.X. designed the research; L.Z. collected the materials and performed genome sequencing; L.H., Y.M., W.Y., T.L., J.J., L.W., and L.F. conducted the genome assembly, annotation and evolution-related data analysis; L.H., Q.H., and J.L. wrote the paper. All authors read and approved the final paper.

Data availability
Raw Illumina-short reads and Hi–C reads used for de novo whole-genome assembly have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive database under accession number PRJNA630530. The genome and related annotation data have been deposited in the National Genomics Data Center (PRJCA002888).

Conflict of interest
The authors declare that they have no conflict of interest.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Information accompanies this paper at (https://doi.org/10.1038/s41438-020-00422-w).

Received: 9 April 2020 Revised: 21 September 2020 Accepted: 30 September 2020
Published online: 01 December 2020

References

1. Schranz, M. E. & Mitchell-Olds, T. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. *Plant Cell.* **18**, 1152–1165 (2006).
2. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).
3. Adams, K. L. & Wendel, J. F. Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8**, 135–141 (2005).
4. Blanc, G. & Wolfe, K. H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell.* **16**, 1667–1678 (2004).
5. Paterson, A. H., Bowers, J. E. & Chapman, B. A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci. USA* **101**, 9903–9908 (2004).
6. Bertioli, D. J. et al. An analysis of synteny of *Arachis* with *Lotus* and *Medicago* sheds new light on the structure, stability and evolution of legume genomes. *BMC Genomics.* **10**, 45 (2009).
7. Tang, H., Bowers, J. E., Wang, X. & Paterson, A. H. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc. Natl Acad. Sci. USA* **107**, 472–477 (2010).
8. Jiao, Y., Li, J., Tang, H. & Paterson, A. H. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell.* **26**, 2792–2802 (2014).
9. Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).
10. Cai, J. et al. The genome sequence of the orchid *Phalaenopsis equestris*. *Nat. Genet.* **47**, 65 (2015).
11. Huang, C. et al. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Mol. Biol. Evol.* **33**, 394–412 (2016).
12. Hegarty, M. J. & Hiscock, S. J. Genomic clues to the evolutionary success of polyploid plants. *Curr. Biol.* **18**, R435–R444 (2008).
13. Sémon, M. & Wolfe, K. H. Reciprocal gene loss between *Tetraodon* and zebrafish after whole genome duplication in their ancestor. *Trends Genet.* **23**, 108–112 (2007).
14. Perumal, S. et al. Elucidating the major hidden genomic components of the A, C, and AC genomes and their influence on *Brassica* evolution. *Sci. Rep.* **7**, 1–12 (2017).
15. Kiefer, M. et al. BrassiBase: introduction to a novel knowledge database on Brassicaceae evolution. *Plant Cell Physiol.* **55**, e3 (2014).
16. Appel, O. & Al-Shehbaz, I. A. *Cruciferae. in Flowering Plants·Dicotyledons* (Springer, 2003).
17. Al-Shehbaz, I. A., Beilstein, M. A. & Kellogg, E. A. Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. *Plant Syst. Evol.* **259**, 89–120 (2006).
18. O'Kane Jr, S. L. Brassicaceae, molecular systematics and evolution of. *Brenner's Encycl. Genet. Second Ed.* 374–376, https://doi.org/10.1016/B978-0-12-374984-0.00169-8 (2013).
19. Blanc, G., Hokamp, K. & Wolfe, K. H. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* **13**, 137–144 (2003).
20. De Bodt, S., Maere, S. & de Peer, Y. Genome duplication and the origin of angiosperms. *Trends Ecol. Evol.* **20**, 591–597 (2005).
21. Soltis, D. E. et al. Polyploidy and angiosperm diversification. *Am. J. Bot.* **96**, 336–348 (2009).
22. Ming, R. et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**, 991–996 (2008).
23. Tang, H. et al. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
24. Fawcett, J. A., Maere, S. & Van de Peer, Y. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc. Natl Acad. Sci. USA* **106**, 5737–5742 (2009).
25. Wang, X. et al. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **43**, 1035 (2011).
26. Kagale, S. et al. Polyploid Evolution of the Brassicaceae during the Cenozoic Era. *Plant Cell.* **26**, 2777–2791 (2014).
27. Guo, X. et al. The genomes of two *Eutrema* species provide insight into plant adaptation to high altitudes. *DNA Res.* **25**, 307–315 (2018).
28. Kang, M. et al. A chromosome-scale genome assembly of *Isatis indigotica*, an important medicinal plant used in traditional Chinese medicine. *Hortic. Res.* **7**, 1–10 (2020).
29. Huang, R. et al. Artificially induced polyploidization in *Lobularia maritima* (L.) Desv. and its effect on morphological traits. *HortScience* **50**, 636–639 (2015).
30. Gómez, J. M. et al. Phenotypic selection and response to selection in *Lobularia maritima*: importance of direct and correlational components of natural selection. *J. Evol. Biol.* **13**, 689–699 (2000).
31. Polunin, O. & Everard, B. *Flowers of Europe: A Field Guide* (Oxford University Press, Oxford, 1969).
32. Yuan, X. Y., Zhang, X. Y., Ma, J. & Hou, X. F. Tissue culture in vitro and establishment of regeneration system of *Lobularia maritima*. *North. Hort.* **8**, 145–146 (2010).
33. Popova, O. V. & Golldack, D. In the halotolerant *Lobularia maritima* (Brassicaceae) salt adaptation correlates with activation of the vacuolar H+-ATPase and the vacuolar Na+/H+ antiporter. *J. Plant Physiol.* **164**, 1278–1288 (2007).
34. Liu, B. et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quant. Biol.* **35**, 62–67 (2013).
35. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
36. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
37. Hunter, S. et al. InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2008).
38. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25 (2000).

Downloaded from academic.oup.com/hr/article/doi/10.1038/s41438-020-00422-w/6445704 by guest on 03 September 2026

39. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
40. Oliver, K. R., McComb, J. A. & Greene, W. K. Transposable elements: powerful contributors to angiosperm evolution and diversity. *Genome Biol. Evol.* **5**, 1886–1901 (2013).
41. Wang, X. et al. Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice. *BMC Bioinforma.* **7**, 447 (2006).
42. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
43. Lysak, M. A., Mandáková, T. & Schranz, M. E. Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. *Curr. Opin. Plant Biol.* **30**, 108–115 (2016).
44. Mandáková, T., Guo, X., Özüdoğru, B., Mummenhoff, K. & Lysak, M. A. Hybridization-facilitated genome merger and repeated chromosome fusion after 8 million years. *Plant J.* **96**, 748–760 (2018).
45. Mandáková, T. & Lysak, M. A. Chromosomal phylogeny and karyotype evolution in $x = 7$ crucifer species (Brassicaceae). *Plant Cell.* **20**, 2559–2570 (2008).
46. Wang, J. et al. Two likely auto-tetraploidization events shaped kiwifruit genome and contributed to establishment of the Actinidiaceae family. *iScience* **7**, 230–240 (2018).
47. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
48. Chandrasekaran, U., Luo, X., Zhou, W. & Shu, K. Multifaceted signaling networks mediated by abscisic acid insensitive 4. *Plant Commun.* **1**, 100040 (2020).
49. Harb, A., Krishnan, A., Ambavaram, M. M. R. & Pereira, A. Molecular and physiological aanalysis of drought stress in *Arabidopsis* reveals early responses leading to acclimation in plant growth. *Plant Physiol.* **154**, 1254–1271 (2010).
50. Kariola, T. et al. Early responsive to dehydration 15, a negative regulator of abscisic acid responses in *Arabidopsis*. *Plant Physiol.* **142**, 1559–1573 (2006).
51. Boudsocq, M. & Sheen, J. CDPKs in immune and stress signaling. *Trends Plant Sci.* **18**, 30–40 (2013).
52. Beilstein, M. A., Al-Shehbaz, I. A. & Kellogg, E. A. Brassicaceae phylogeny and trichome evolution. *Am. J. Bot.* **93**, 607–619 (2006).
53. Guo, X. et al. Plastome phylogeny and early diversification of Brassicaceae. *BMC Genomics.* **18**, 176 (2017).
54. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
55. Dassanayake, M. et al. The genome of the extremophile crucifer *Thellungiella parvula*. *Nat. Genet.* **43**, 913–918 (2011).
56. Sudmant, P. H. et al. Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
57. Arnaiz, A. et al. *Arabidopsis* kunitz trypsin inhibitors in defense against spider mites. *Front. Plant Sci.* **9**, 986 (2018).
58. Zschiesche, W. et al. The zinc-binding nuclear protein HIPP3 acts as an upstream regulator of the salicylate-dependent plant immunity pathway and of flowering time in *Arabidopsis thaliana*. *N. Phytol.* **207**, 1084–1096 (2015).
59. Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–113 (1970).
60. Azevedo, C. et al. Role of SGT1 in resistance protein accumulation in plant immunity. *EMBO J.* **25**, 2007–2016 (2006).
61. Tör, M. et al. *Arabidopsis* SGT1b is required for defense signaling conferred by several downy mildew resistance genes. *Plant Cell.* **14**, 993–1003 (2002).
62. Holt, B. F. Antagonistic control of disease resistance protein stability in the plant immune system. *Science* **309**, 929–932 (2005).
63. Austin, M. J. Regulatory role of SGT1 in early R gene-mediated plant defenses. *Science* **295**, 2077–2080 (2002).
64. Cheung, M.-Y. et al. ATP binding by the P-loop NTPase OsYchF1 (an unconventional G protein) contributes to biotic but not abiotic stress responses. *Proc. Natl Acad. Sci. USA* **113**, 2648–2653 (2016).
65. Pascuan, C., Frare, R., Alleva, K., Ayub, N. D. & Soto, G. mRNA biogenesis-related helicase eIF4AIII from *Arabidopsis thaliana* is an important factor for abiotic stress adaptation. *Plant Cell Rep.* **35**, 1205–1208 (2016).
66. Postlethwait, J. H. et al. Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res.* **10**, 1890–1902 (2000).
67. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708 (1997).
68. Blanc, G., Barakat, A., Guyot, R., Cooke, R. & Delseny, M. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell.* **12**, 1093–1101 (2000).
69. Vision, T. J., Brown, D. G. & Tanksley, S. D. The origins of genomic duplications in *Arabidopsis*. *Science* **290**, 2114–2117 (2000).
70. Simillion, C., Vandepoele, K., Van Montagu, M. C. E., Zabeau, M. & Van de Peer, Y. The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **99**, 13627–13632 (2002).
71. Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438 (2003).
72. Doyle, J. J. et al. Evolutionary genetics of genome merger and doubling in plants. *Annu. Rev. Genet.* **42**, 443–461 (2008).
73. Madlung, A. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity* **110**, 99 (2013).
74. Ramsey, J. Polyploidy and ecological adaptation in wild yarrow. *Proc. Natl Acad. Sci. USA* **108**, 7096–7101 (2011).
75. Diallo, A. M., Nielsen, L. R., Kjær, E. D., Petersen, K. K. & Ræbild, A. Polyploidy can confer superiority to West African *Acacia senegal* (L.) Willd. trees. *Front. Plant Sci.* **7**, 821 (2016).
76. Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Res.* **24**, 1334–1347 (2014).
77. Vanneste, K., Maere, S. & Van de Peer, Y. Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, 20130353 (2014).
78. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15 (1987).
79. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
80. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
81. Wu, H. et al. A high-quality *Actinidia chinensis* (kiwifruit) genome. *Hortic. Res.* **6**, 1–9 (2019).
82. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
83. Li, H. BFC: correcting Illumina sequencing errors. *Bioinformatics* **31**, 2885–2887 (2015).
84. Xu, H. et al. FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS ONE* **7**, e52249 (2012).
85. Kajitani, R. et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).
86. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
87. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
88. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. In *Current Protocols in Bioinformatics* (John Wiley & Sons, Inc., 2002).
89. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
90. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* **9**, 18 (2008).
91. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
92. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
93. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
94. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
95. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).

Downloaded from academic.oup.com/hr/article/doi/10.1038/s41438-020-00422-w/6445704 by guest on 03 September 2026

96. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
97. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, 1–22 (2008).
98. Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
99. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
100. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
101. Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971–2972 (2006).
102. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
103. Lechner, M. et al. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinform.* **12**, 124 (2011).
104. Sahm, A., Bens, M., Platzer, M. & Szafranski, K. PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes. *Nucleic Acids Res.* **45**, e100 (2017).

Downloaded from academic.oup.com/hr/article/doi/10.1038/s41438-020-00422-w/6445704 by guest on 03 September 2026