


Resource Article: Genomes Explored

# A chromosome-level genome assembly of an alpine plant *Crucihimalaya lasiocarpa* provides insights into high-altitude adaptation

Landi Feng, Hao Lin, Minghui Kang, Yumeng Ren, Xi Yu, Zhanpeng Xu, Shuo Wang, Ting Li, Wenjie Yang, and Quanjun Hu  \*

Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, China

\*To whom correspondence should be addressed. Tel. +86 132 5819 5631. Fax. +028 8541 2053. Email: huquanjun@scu.edu.cn

Received 9 November 2021; Editorial decision 16 January 2022

## Abstract

It remains largely unknown how plants adapt to high-altitude habitats. *Crucihimalaya* (Brassicaceae) is an alpine genus occurring in the Qinghai–Tibet Plateau characterized by cold temperatures and strong ultraviolet radiation. Here, we generated a chromosome-level genome for *C. lasiocarpa* with a total size of 255.8 Mb and a scaffold N50 size of 31.9 Mb. We first examined the karyotype origin of this species and found that the karyotype of five chromosomes resembled the ancestral karyotype of the Brassicaceae family, while the other three showed strong chromosomal structural variations. In combination with the rough genome sequence of another congener (*C. himalaica*), we found that the significantly expanded gene families and positively selected genes involved in alpine adaptation have occurred since the origin of this genus. Our new findings provide valuable information for the chromosomal karyotype evolution of Brassicaceae and investigations of high-altitude environment adaptation of the genus.

**Key words:** *Crucihimalaya*, adaptation, *de novo* genome, karyotype evolution, high altitude

## 1. Introduction

Alpine plants occurring in high-altitude habitats with cold temperatures and strong ultraviolet (UV) radiation have evolved special adaptations to withstand these abiotic stresses.<sup>1–3</sup> Genomic studies of non-model alpine species have provided great insights into the genetic mechanisms underlying such adaptations. For example, genomic comparisons of two *Eutrema* (Brassicaceae) species distributed in the high or the low altitude<sup>4</sup> suggest that gene family expansions and duplicated genes in the alpine species are involved in their disease resistance, DNA damage repair, reproduction and cold tolerance. Further population genomic analyses of the other species with contrasted altitudinal distributions also identified the positively selected genes related to high-altitude adaptation in multiple genera.<sup>5–10</sup>

Although the revealed genes vary greatly, their functions are always annotated to be involved in similar molecular pathways, including both DNA repairs and other physiological adaptations in response to alpine habitats.

The genus *Crucihimalaya* (belongs to Brassicaceae) contains 2–4 self-pollination diploid species with chromosome number of  $2n = 16$ , mainly distributed in the high-altitude regions of the Qinghai–Tibet Plateau.<sup>11,12</sup> The species in this genus were previously placed in the genera *Sisymbrium* or *Arabidopsis* because of their morphological similarity until the recent taxonomic revisions.<sup>11</sup> Phylogenetic analyses based on sequence variations of the nuclear ribosomal internal transcribed spacer (ITS) region and other nuclear genes<sup>13</sup> further supported the separation of this genus from *Sisymbrium* or

*Arabidopsis*. The rough genome of one species from this genus, *C. himalaica*, was reported based on the Second-Generation Sequencer.<sup>14</sup> Compared with the low-altitude genera *Capsella* and *Arabidopsis*, gene families related to disease resistance in this species were found to have significantly contracted while those to DNA repair and ubiquitin-mediated proteolysis expanded. In addition, genes related to reproductive processes were found to have experienced positive selection and/or functional loss in *C. himalaica*, although it remains unclear whether the identified genomic changes are common for all species of the genus occurring in the high altitudes or especially for *C. himalaica*. In addition, as the genome sequences of this species had not been assembled into chromosomes, the karyotype origin of the genus *Crucihimalaya* remains unresolved. Although the karyotype evolution in Brassicaceae is highly diverse, it can be traced to the recombination of two basic ancestral karyotypes.<sup>15,16</sup> The chromosome-level genome sequence could be used to examine the karyotype origin of the genus from such ancestral karyotypes.<sup>17</sup>

In this study, we report a chromosome-level reference genome for another *Crucihimalaya* species, *C. lasiocarpa*, with a total size of 255.8 Mb and a scaffold N50 size of 31.9 Mb. Almost all protein-coding genes (99.13%) are anchored on eight chromosomes. In addition, we first examined the karyotype origin for the genus based on this chromosome-level genome sequence. Two other *Crucihimalaya* species without genomic resources are very similar to *C. lasiocarpa* and *C. himalaica*<sup>11,12</sup> and these two species with available genomes likely represent the genetic diversity of the genus. Therefore, we further examined genomic evolution of the genus and how it was associated with alpine adaptation after conducting comparative genomic analyses of the two *Crucihimalaya* species. The renewed high-quality genome sequence for this genus will be highly useful for exploring the relative contributions of various genomic mechanisms in driving the adaptive evolution to high-altitude environments.

## 2. Materials and methods

### 2.1. Sample collection, DNA extraction and sequencing

Seeds of *C. lasiocarpa* were collected from Tibet, China (altitude 4,000 m, N 39.718°, E 91.120°) and germinated in the greenhouse (Fig. 1a). The whole steps of library construction and sequencing were performed at Grandomics Biotechnology Co., Ltd (Wuhan, China). The high-quality genomic DNA was extracted from fresh leaves of single individual by the Qiagen Genomic kit. Size selection was performed using the Blue-pippin system (Sage Science) to obtain DNA fragments from 10 to 50 kbp. The following library constructed using 1D ligation kit (SQK-LSK109) were sequenced on a PromethION platform (Oxford Nanopore Technologies, UK). Base calling was completed by Guppy v3.2.2 18 with default parameters. ONT reads with mean quality scores  $\geq 7$  (q7) were retained for the following genome assembly.

Furthermore, the paired-end Illumina libraries with insert sizes of 400 bp were prepared using the Illumina Genomic DNA Sample Preparation Kit, which were then sequenced on an Illumina NovaSeq platform (Beijing, Grandomics Biotechnology Co., Ltd). The filtered clean data was used for *k*-mer analysis and also the error correction. Furthermore, RNA-Seq reads of fresh leaf and root were generated for gene annotation using the MGI-2000 platform (Shenzhen, BGI Genomics Co., Ltd).

For Hi-C sequencing, fresh leaves from one plant were first fixed by formaldehyde. The fixed chromatin was then digested with DpnII restriction endonuclease.<sup>18</sup> The digested nucleotide fragments

were biotinylated with further proximity ligation to form chimeric circles.<sup>19,20</sup> The DNA was further purified and sheared into fragments again for preparing the sequencing library. Finally, the chromatin conformation capture library was sequenced on an Illumina NovaSeq PE150 platform.

### 2.2. Genome assembly and chromosome construction

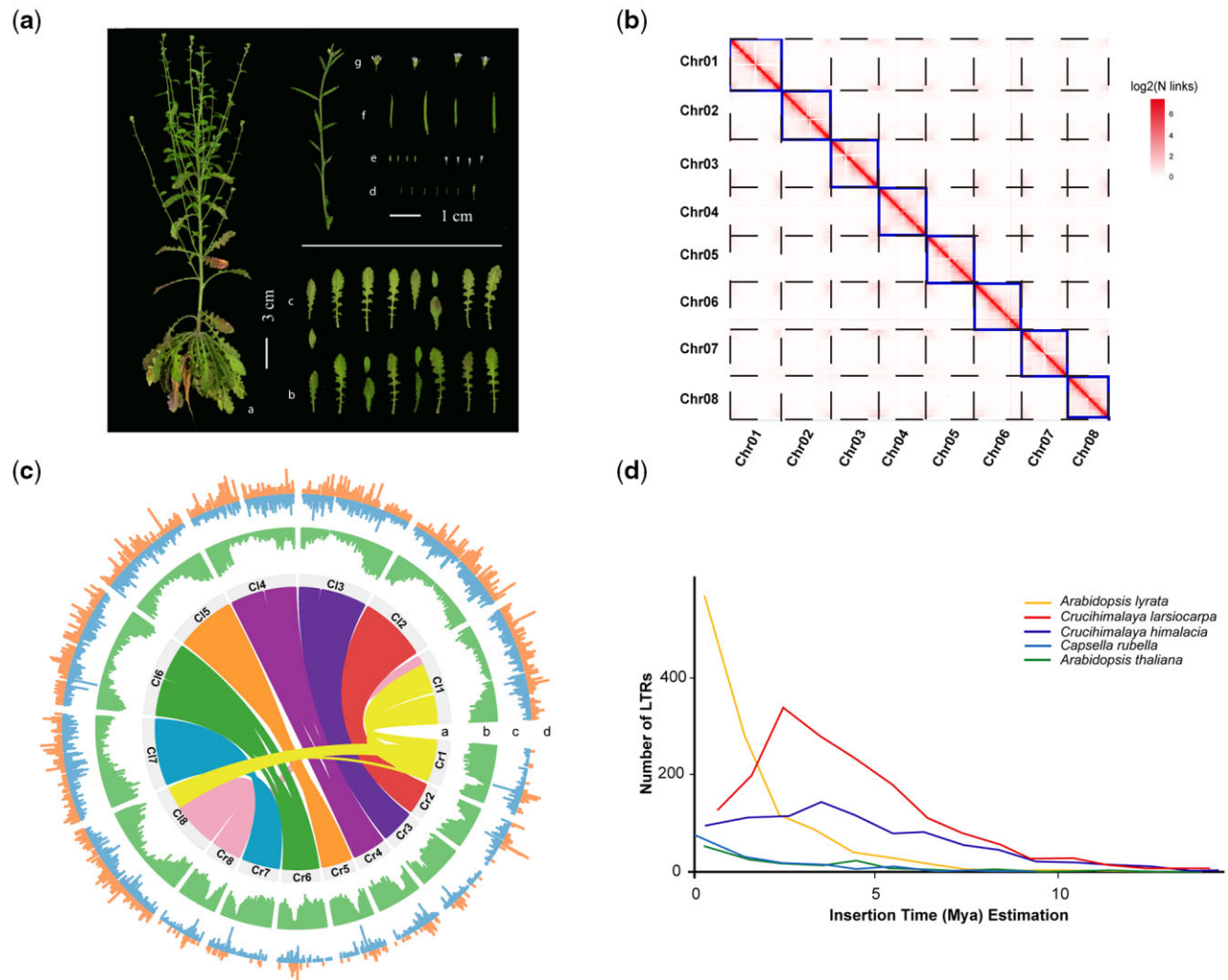
Before genome assembly, *k*-mer frequency distribution analysis was applied to estimate genome size and heterozygosity (size 21 bp in Illumina DNA short reads). A total of 37.72 Gb raw Nanopore long reads were corrected, assembled using NextDenovo v2.0-beta.1 (<https://github.com/Nextomics/NextDenovo> (24 January 2022, date last accessed)) with parameter of seed\_cutoff = 38k, reads\_cutoff = 1k. The preliminary assembly of *C. lasiocarpa* was followed by three rounds of polishing using NextPolish<sup>21</sup> with both nanopore and short Illumina sequencing reads. Then, we mapped Hi-C data to the contigs using Juicer v1.6.2 pipeline<sup>22,23</sup> and built primary scaffolds by the 3D-DNA v180922 with default parameters.<sup>24</sup> Juicebox Assembly Tools v1.9.8 was used to visualize and manually curate the assembly.<sup>25</sup> Afterwards, we processed another round of scaffolding by 3D-DNA v180922 to obtain the final pseudochromosomes. To assess the accuracy of the genome assembly, Illumina short reads were mapped against the genome using BWA v0.7.12-r1039 with default parameters.<sup>26</sup> Benchmarking Universal Single-Copy Orthologous gene analysis (BUSCO) with the gene content of Embryophyta\_odb10 were used to further evaluate the completeness of the assembled genome.<sup>27</sup> LAI (LTR Assembly Index) scores were calculated using the LTR\_retriever pipeline.<sup>28</sup>

### 2.3. Repeat annotation

Repetitive elements in *C. lasiocarpa* genome were identified by a combination method of homology-based and *de novo* strategies using RepeatModeler v1.0.11<sup>29</sup> and RepeatMasker v4.0.7<sup>30</sup> with default settings. LTRharvest v1.5.10<sup>31</sup> and LTRFinder v1.06<sup>32</sup> were used to detect full-length long terminal repeat retrotransposons (FL-LTR-RTs) in genome. The resulting outputs (.scn) from both analyses were fed into the LTR\_retriever v1.9 programme<sup>28</sup> to extract FL-LTR-RTs. The insertion time (T) for each LTR retrotransposons was calculated by the formula:  $T = K/2r$  using a substitution rate (*r*) of  $7 \times 10^{-9}$  substitutions per site per year, where *K* represented genetic distance.<sup>33</sup>

### 2.4. Gene prediction and function annotation

For accurate prediction of genes in the *C. lasiocarpa* genome, an integrated computational approach was adopted based on transcript mapping, homologous protein alignment and *ab initio* prediction. Illumina RNA-Seq raw reads were first processed using Trimmomatic v0.33<sup>34</sup> to detect and remove adapter and low-quality bases (Phred quality score <20). The transcriptome assembly was then performed by Trinity v2.6.6 using filtered reads.<sup>35</sup> The assembled transcripts were further aligned to the assembled genome to carry out ORF prediction by PASA v2.1.0 pipeline.<sup>36</sup> Protein sequences from *Aethionema rabicum*, *A. lyrata*, *Arabidopsis thaliana*, *Boechera stricta*, *Brassica rapa*, *Capsella grandiflora*, *Capsella rubella*, *Carica papaya*, *Leavenworthia alabamica*, and *Tarenaya hasleriana* (Supplementary Table S1) were aligned to the *C. lasiocarpa* genome using exonerate v2.4.0.<sup>37</sup> *Ab initio* gene prediction was carried out by Augustus v3.2.3 and GlimmerHMM v3.0.452 with parameter files generated by PASA self-trained gene models.<sup>38</sup> Gene



**Figure 1.** Summary of *Crucihimalaya lasiocarpa* genome assembly. (a) Photo of *C. lasiocarpa*: schematic representations of (i) aerial parts, (ii) adaxial surfaces of leaves, (iii) abaxial surfaces of leaves, (iv) stamens, (v) calyx and petal, (vi) fruit pods and (vii) flowers. (b) The Hi-C chromatin interaction map for the eight pseudochromosomes of *C. lasiocarpa*. (c) Genome comparison between *C. lasiocarpa* and *Cap. rubella*: (i) syntentic relationships between *C. lasiocarpa* and *Cap. rubella* genomes; (ii) gene density (window size = 100 kb, nonoverlapping); (iii) density distribution of *Copia* elements (window size = 100 kb, non-overlapping); and (iv) density distribution of *Gypsy* elements (window size = 100 kb, non-overlapping). (d) Number distribution and estimated insertion times of intact LTR retrotransposons.

model evidence from the aforementioned results were integrated into a non-redundant set of gene annotations with EvidenceModeler v1.1.1.<sup>39</sup> All predicted genes were analysed for functional domains and homologs by searching against the InterPro v32 using InterProScan<sup>40</sup> and alignment to the integrated protein sequence databases including Swiss-Prot and TrEMBL.<sup>41</sup> Subsequently, GO and KO annotation was analysed based on Blast2GO v2.5<sup>42</sup> and KEGG pathway database.<sup>43</sup>

## 2.5. Genomic blocks identification

To better understand the evolutionary scenario of the *C. lasiocarpa* genome, we identified syntentic blocks in the *C. lasiocarpa* genome relative to the 22 genomic blocks (A–X) in ancestral crucifer karyotype (ACK).<sup>15,44</sup> As genomic blocks in the ACK were defined using the *A. thaliana* gene IDs as start and end coordinates, homologous gene pairs were detected by BLASTP with a cut-off *e* value of  $1e^{-5}$  of the predicted protein sequences against the *A. thaliana* proteome.

Next, syntentic relationships were further determined using MScanX<sup>45</sup> and LAST v.946.<sup>46</sup> Based on the collinearity in each genomic block against *A. thaliana*, we determined the corresponding boundaries and intervals of each block in *C. lasiocarpa* and renamed the pseudochromosomes. The genomic blocks in *Capsella rubella* were obtained through the same method applied in *C. lasiocarpa*. In order to eliminate false chromosomal rearrangements caused by assembly errors, the raw Hi-C reads of *C. lasiocarpa* were used to map to the whole genomic sequence of *Capsella rubella* using 3D-DNA v180922. Finally, we examined the order, orientation and contiguity of genes across potential adjacent segments, where inter-chromosomal rearrangements were inferred. The most parsimonious scenario of the fusion and fission events during genome evolution of *C. lasiocarpa* from its ancestral chromosomes was determined. The breakpoint regions were identified by pairwise genome alignment using LAST and MScanX. By inferring putative homologous genes and collinear genes between *Capsella rubella* and *C. lasiocarpa*, we drew a homologous gene dot plot using WGDI.<sup>47</sup>

## 2.6. Phylogenetic and gene family analyses

Orthologous groups were identified using OrthoFinder v2.3.12<sup>48</sup> by all-versus-all BLASTP alignments (E-value  $\leq 1e^{-5}$ ) with protein sequences from *C. lasiocarpa*, *Crucihimalaya himalaica*, *Capsella rubella*, *A. thaliana*, *Eutrema heterophyllum*, *Eutrema yunnanense*, *Eutrema salsugineum* and *Aethionema arabicum* (Supplementary Table S2). We kept the longest transcripts of each gene model to eliminate redundancy caused by alternative splicing variations. Orthogroups with only one gene copy per species (Single-copy orthogroups) were collected, and aligned using Mafft v7.313 with globalpair G-INS-i strategy.<sup>49</sup> The alignments of each single copy orthogroups were concatenated into a super-alignment. The super-alignments were then filtered by Gblocks v.0.91b<sup>50</sup> to remove gap regions. Subsequently, phylogenetic trees were constructed by RAxML v8.2.11<sup>51</sup> using the GTRGAMMA model and performed 100 bootstrap analyses to test the robustness of each branch.

Divergence time estimation was performed by MCMCTree in PAML v4.9 package,<sup>52</sup> which implements the Markov chain Monte Carlo algorithms of Yang and Rannala.<sup>53</sup> Analyses were run for 100,000 generations with a burn-in of 1,000 iterations. The calibrated molecular clock used to estimate divergence time (32–43 MYA between the *A. arabicum* and *A. thaliana*) was obtained from the TimeTree database.<sup>54</sup> All MCMCTree calculations were run twice to ensure convergence.

We employed CAFE v4.2<sup>55</sup> based on a Bayesian method to discover gene family contraction and expansion events using OrthoFinder results as input. The rooted and bifurcating tree from the phylogenetic analysis were used to time-calibrate the gene trees. To check the significance of contraction/expansion events at specific branches, we computed the branch-specific *P*-value and family-wide *P*-value with the Viterbi method for each orthogroup. Genes in significantly expanded families were then used for Gene Ontology enrichment analysis. The KOBAS software<sup>56</sup> was also used to test the statistical enrichment of genes in KEGG pathways.

## 2.7. Identification of positively selected genes

To search for genes that evolved under positive selection (PSGs), single-copy gene families were extracted by OrthoFinder. Based on the non-synonymous to synonymous substitution ratio, the Branch-Site Model and Branch Model in CODEML from PAML package were used to detect selection with the ancestral branch leading to the *C. lasiocarpa* and *C. himalaica* set as foreground branch. For the Branch-Site Model, the  $\chi^2$  test was conducted for each orthologous to assess statistical significance of Model A and Model A null. Finally, Bayes empirical Bayes (BEB) analysis<sup>57</sup> was used to identify gene with positively selected sites of posterior probabilities greater than or equal to 0.99. For Branch Model, we ran the one-ratio branch model (null model which assumes that all branches have been evolving at the same rate) and the multi-ratio model (alternative model which supposing foreground branch to evolve under a different rate) to estimate the *dN/dS* ratio in each orthologs. A likelihood ratio test (LRT) with *df* = 1 was employed to discriminate between the null model and the alternative model.<sup>58</sup> Genes with a *P*-value < 0.05 and a higher *dN/dS* value in the foreground than the background branches were regarded as positively selected genes in the foreground.<sup>59</sup>

## 2.8. S-locus structure and self-fertilization

Known that loss of function at the S-locus is the reason for self-fertilization of *C. himalaica* and the genus *Crucihimalaya* is self-

compatible,<sup>14</sup> we searched S-haplogroups sequences in the *C. lasiocarpa* to explore if same mutations could be in *C. lasiocarpa*. The published S-haplogroups sequences of *Crucihimalaya himalaica*, *Capsella grandiflora*, *A. thaliana*, *Arabidopsis halleri*, *A. lyrata* and *B. rapa* were retrieved from NCBI and used as query in BLASTP searches against the genome assembly of *C. lasiocarpa*. Via searches, we manually annotated the candidate homologous gene of S locus SCR gene and the flanking genes U-box gene. Intriguingly, there was a candidate protein hit, which matched ARK3 and SRK. Due to the confusing role of candidate ARK3 and SRK protein, the SRK and ARK3 protein of *Arabidopsis* were downloaded and aligned by MAFFT. Then, a maximum likelihood tree both including SRK and ARK3 was constructed using RaxML. All candidate proteins of S-locus were further confirmed by hmmsearch against the Pfam database.

## 3. Results

### 3.1. Genome assembly and gene annotation

The genome size of *C. lasiocarpa* is estimated at ~253.6 Mb, with a heterozygosity rate of 0.04% based on *k*-mer analysis (Supplementary Fig. S1). To obtain a high-quality, chromosome-scale genome assembly, we produced 139× coverage of Oxford Nanopore long-reads sequencing data (37.72 Gb), 53× coverage of paired-end Illumina short reads sequencing data (18.00 Gb) and 104× coverage of paired-end Hi-C reads (54.99 Gb) (Supplementary Tables S3 and S4). Firstly, the Oxford Nanopore long reads and Illumina short reads were used for primary assembly, comparison and error correction, which finally produced a total of 255.82 Mb genome assembly with an average contig N50 of 14.98 Mb, close to the estimated genome size of 253.6 Mb (Supplementary Fig. S1). With the assistance of Hi-C reads, we then anchored contigs into eight pseudo-chromosomes using 3D-DNA pipeline (Figure 1b). Compared to the previously published *C. himalaica* genome assembly based on Illumina short-read technology that produced an assembly of 583 scaffolds (3,983 contigs), with a scaffold N50 of just 2.0 Mb (contig N50 = 136.3 Kb), the final genome assembly of *C. lasiocarpa* exhibited a total size of 255.81 Mb, including 20 scaffolds with a 31.9 Mb scaffold N50 size, and the largest scaffold size was 35.0 Mb. The gap rate of the assembled *C. lasiocarpa* sequences was estimated to be only 0.003%, while it was 1.63% for the *C. himalaica* genome assembly (Table 1). We further evaluated the completeness of the assembled genome, and found high completeness (99.6% of *C. lasiocarpa* and 99.7% of *C. himalaica*) rate of both assemblies as evidenced by BUSCO analysis<sup>60</sup> (Supplementary Fig. S2). The long terminal repeat (LTR) Assembly Index (LAI), which evaluates the contiguity of intergenic and repetitive regions of genome assemblies based on the intactness of LTR retrotransposons (LTR-RTs),<sup>61</sup> was 17.64 for *C. lasiocarpa* genome assembly, which was substantially higher than the LAI values obtained for the *C. himalaica* genome and other relatives under comparison (Supplementary Fig. S3).

Based on a combination of *de novo* prediction, homology searching, and transcriptome-based approaches, we annotated 24,169 protein-coding genes in the *C. lasiocarpa* genome. Around 99.13% of the genes (23,960 out of 24,169) were anchored to eight chromosomes while only 0.86% (208 out of 24,169) were left on the unanchored scaffolds. The average gene length, coding sequence length and an average exon number were estimated to be 2,582 base pairs (bp), 236 bp and 5.53 exons, respectively (Supplementary Fig. S4). These protein-coding genes were further functionally annotated

**Table 1.** Statistics for *C. lasiocarpa* and *C. himalaica* genome assemblies.

Species	<i>Crucihimalaya lasiocarpa</i>	<i>Crucihimalaya himalaica</i>
Sequencing platform	ONT PromethION	Illumina HiSeq 2500
Assembly size (bp)	255,812,582	234,722,603
GC %	36.49	36.38
Number of scaffolds	20	583
Longest scaffold (bp)	35,013,560	8,343,586
Scaffold N50 size (bp)	31,983,042	2,088,603
Scaffold N90 size (bp)	27,759,296	470,087
Number of Scaffold N50	4	34
Number of Scaffold N90	8	129
Number of contigs	58	3,983
Longest contig (bp)	21,500,254	1,756,581
Contig N50 size (bp)	14,980,479	136,392
Contig N90 size (bp)	11,549,935	32,421
Number of Contig N50	8	406
Number of Contig N90	15	1711
Gap %	0.003	1.63
Number of genes	24,169	27,019

based on Swiss-Prot, InterPro, GO and KEGG Pathway databases (Supplementary Table S5). The average gene length, coding sequence length and an average exon number were estimated to be 2,582 base pairs (bp), 236 bp and 5.53 exons, respectively (Supplementary Fig. S3). In contrast to the many short-length genes identified in *C. himalaica*,<sup>14</sup> the gene lengths of *C. lasiocarpa* are normally distributed as those of *A. thaliana*. Furthermore, the completeness of the gene annotation was also assessed using BUSCO, and the result revealed a higher proportion of complete single-copy orthologs in the genome assembly of *C. lasiocarpa* (97.2%) compared to that in *C. himalaica* (96%) (Supplementary Table S6), suggesting that *C. lasiocarpa* genome seems to have a high gene-annotation quality.

### 3.2. Evolutionary origin of repeat elements in *Crucihimalaya lasiocarpa*

Based on *de novo* and homology prediction approaches, a total of 134.5 Mb (52.58%) of the assembled *C. lasiocarpa* genome was identified as repeat regions with LTR elements being identified as the major class (27.01%) (Supplementary Table S7). The proportion of repetitive elements near the centromere is higher than other regions (Fig. 1c). *Crucihimalaya lasiocarpa* contains a higher proportion of repeat elements than *C. himalacia*, *A. thaliana* and *Capsella rubella*.<sup>62,63</sup> We did not find that *C. lasiocarpa* undergoes an additional species-specific WGD event compared with *C. lasiocarpa*, *A. thaliana* and *Capsella rubella* based on the distribution of synonymous substitutions per synonymous site (Ks) among paralogous genes within the genome (Supplementary Fig. S5).

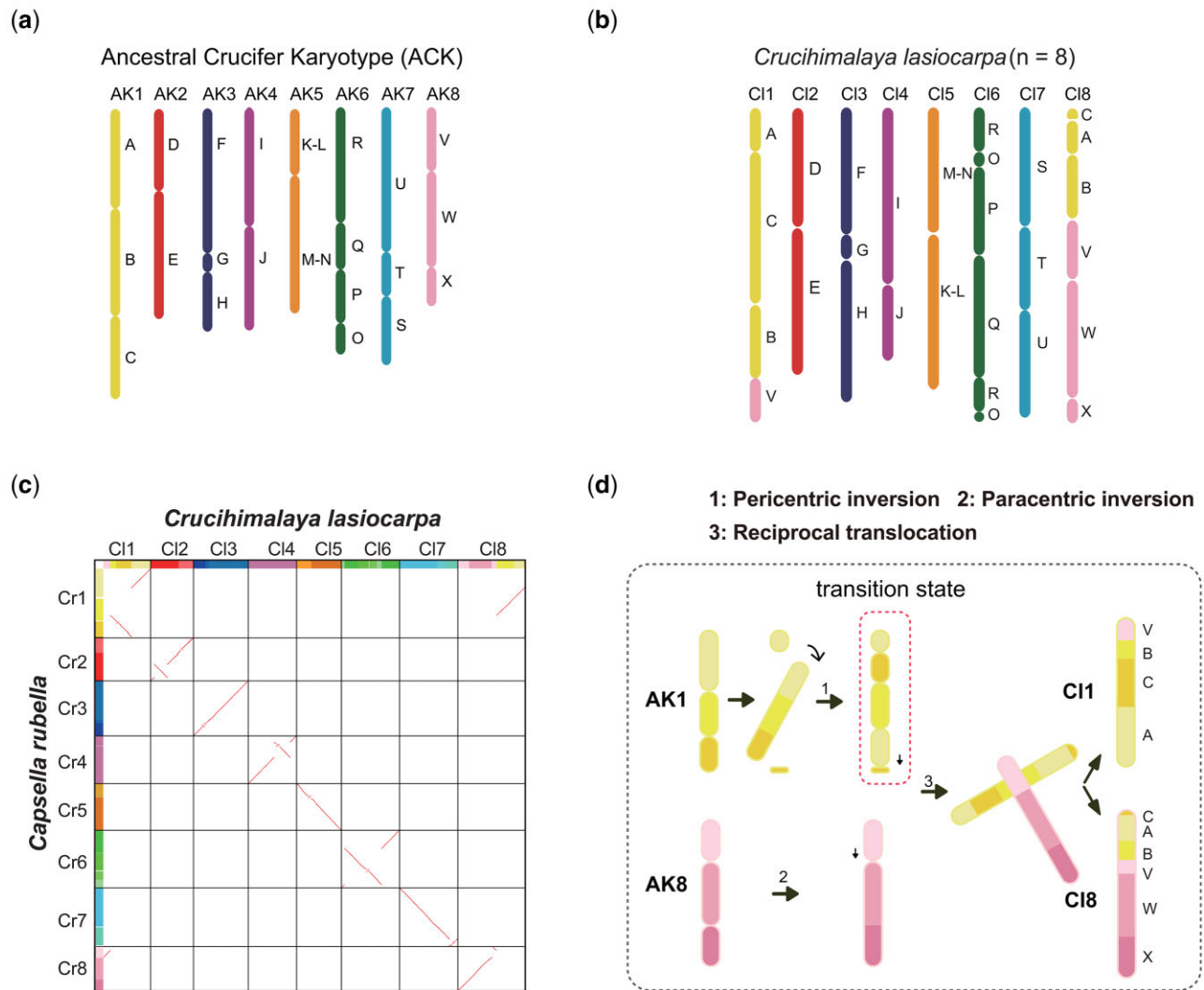
We further explored the evolutionary dynamics of the intact long terminal repeat retrotransposons (LTR-RTs) in the genomes of *C. lasiocarpa* and other three closely related species. A total of 1,719,930,178 and 1,166 LTR-RTs were identified in *C. lasiocarpa*, *C. himalacia*, *Capsella rubella* and *A. lyrata* genomes, respectively. LTR retrotransposons in the genome of *C. lasiocarpa* had recently undergone a rapid proliferation ~2.5 million years ago (Fig. 1d). As the insertion of an LTR into a new position in the genome might lead to alternative splicing of a particular transcript through various mechanisms, resulting in new genetic and phenotypic variations with potential adaptive significance.<sup>64</sup> Compared with *C. himalacia*, we

found that 152 orthologous genes in *C. lasiocarpa* contained specific insertion of *Gypsy* or *Copia* element. Further analysis revealed that these genes exhibited significantly higher protein evolutionary rates (*Ka/Ks*) compared with the genomic background (Supplementary Fig. S6).

### 3.3. Karyotype origin of *Crucihimalaya lasiocarpa*

Chromosomal structural variations play a critical role in phenotypic variation and environmental adaptation.<sup>65</sup> Using comparative chromosome painting (CCP) techniques, Ancestral Crucifer Karyotype (ACK) for the family Brassicaceae with eight chromosomes ( $n=8$ ) with 22 conserved genomic blocks (GBs, A to X) were suggested<sup>66</sup> (Fig. 2a). These defined GBs and the changed compositions from the ACK were used to examine karyotype evolution in other genera and species. As the karyotype of *Capsella rubella* was suggested to be similar to the ACK of the family Brassicaceae,<sup>63</sup> we compared karyotypes of *C. lasiocarpa* and *Capsella rubella* using previously reported methods (Supplementary Figs S7 and S8).<sup>17,67</sup> We compared the genomic blocks of *C. lasiocarpa* with inferred ancestral karyotypes, including the ACK ( $n=8$ ), proto-calepineae karyotype (PCK;  $n=7$ ) and translocated PCK (tPCK;  $n=7$ ). The karyotype of *C. lasiocarpa* was inferred to evolve from the ACK. The *C. lasiocarpa* karyotype comprises six relatively conserved chromosomes (CL2, 3, 4, 5, 6 and 7) and two chromosomes with structural variations occurring (reciprocal translocation and inversion) (Fig. 2b).

The inversion fragments in CL2, 3, 4, 5 and 7 chromosomes are small and do not span GBs boundary, so the order between the GBs in the chromosomes remained unchanged. CL6 possesses a pericentric inversion with a length of ~25M. This inversion spans the four GBs of CL6 (R, Q, P and O) as the longest inversion across the whole genome (Supplementary Fig. S9). The Hi-C interaction heatmap showed that this chromosomal rearrangement was not caused by assembly errors (Supplementary Fig. S9). The majority of these events occurred in the pericentromeric regions of the *C. lasiocarpa* chromosomes (Fig. 2c and d).



**Figure 2.** Ancestral crucifer karyotype and *Crucihimalaya lasiocarpa* genomic blocks. (a) The ancestral genomes ACK comprising 22 ancestral GBs. (b) Twenty-two GBs and their positions within the *C. lasiocarpa* genome. (c) Syntenic dot plot of *C. lasiocarpa* and *Cap. rubella*. The dot plot was generated using programme WGDI. Assignment to genomic blocks is given on the left for *Capsella* and above for *Crucihimalaya*. Syntenic genes are coloured by Ks values. Only gene pairs with Ks value lower than 0.6 are retained. (d) Chromosomal rearrangements illustrating the origin of *Crucihimalaya* genome ( $n=8$ ) from ACK-like genome ( $n=8$ ) are showed.

### 3.4. Phylogenetic analyses and expansion and contraction of gene families

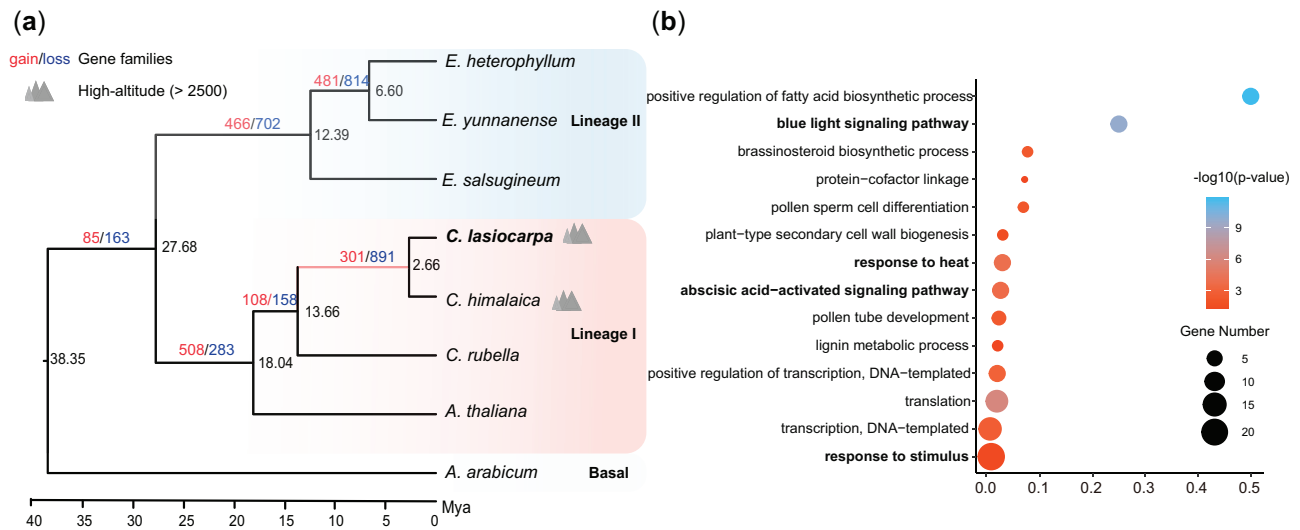
A total 5,993 single-copy gene families were used to construct a maximum-likelihood phylogenetic tree for *C. lasiocarpa* and closely related species (Fig. 3a). The phylogenetic tree showed that the *C. lasiocarpa* was sister to *C. himalaica*, which diverged about 2.66 million years ago (MYA). The *C. lasiocarpa*-*C. himalaica* clade diverged from *Capsella rubella* ~13.66 MYA.

Adaptation is greatly favoured by origination of new genes.<sup>68</sup> Gene duplication offers a rapid way to produce new genes.<sup>69</sup> We found 301 gene families expanded and 891 gene families contracted in the clade comprising two alpine species of the genus *Crucihimalaya*, *C. lasiocarpa* and *C. himalaica* (Fig. 3a). A total of 56 gene families were significantly expanded ( $P$ -value < 0.05), which are enriched for 'blue light signalling pathway', 'response to stimulus' and 'abscisic acid-activated signalling pathway' (Fig. 3b). These pathways have long been proven to be associated with resisting

strong UV radiation and tolerating low temperature.<sup>70</sup> We also found that some ubiquitin-conjugating gene families expanded in these species (Supplementary Table S8). Some gene families found functionally involved in biotic stress significantly contracted, including those related to camalexin biosynthesis, farnesene biosynthesis and Indole-3-acetate biosynthesis (Supplementary Table S9) which were found as special phytoalexin in response to bacterial and fungal pathogens<sup>71,72</sup> and defending against nematodes.<sup>73</sup>

### 3.5. Identification of candidate genes related to High-Altitude adaptation of the genus

Stresses may lead to positive selections of the related genes.<sup>74,75</sup> We identified positively selected genes (PSGs) for the clade comprising *C. lasiocarpa* and *C. himalaica*. Using Branch Model and Branch-Site Model of codeml in the PAML package, we totally identified 403 PSGs (Supplementary Table S10). The functions of the significantly



**Figure 3.** Phylogenetic analysis of the *Crucihimalaya lasiocarpa* genome. (a) The phylogenetic placement of *C. lasiocarpa*, divergence time (million years ago, MYA, black), gene family expansions (red) and contractions (blue) are displayed on a maximum likelihood (ML) tree constructed from 5,993 shared single-copy gene families. (b) Gene ontology (GO) enrichment of significantly expanded gene family of high-altitude clade (highlighted by pink) in a. The colour of circles represents the statistical significance of enriched GO terms. The size of the circles represents the number of genes in a GO term.

positively selected genes (PSGs) were associated with stress tolerance and alpine survival (Fig. 4 and Supplementary Table S10). For example, the gene *HOS15* was found to be involved in resisting cold stress in plants through mediating deacetylation of histone.<sup>76</sup> Both genes *MLH3* and *MSH1* play an important role in repairing DNA mismatches and correcting insertion-deletion loops,<sup>73</sup> while *RLK7* is an important transcription factor to regulate resistances to abiotic stress in plants.<sup>73</sup>

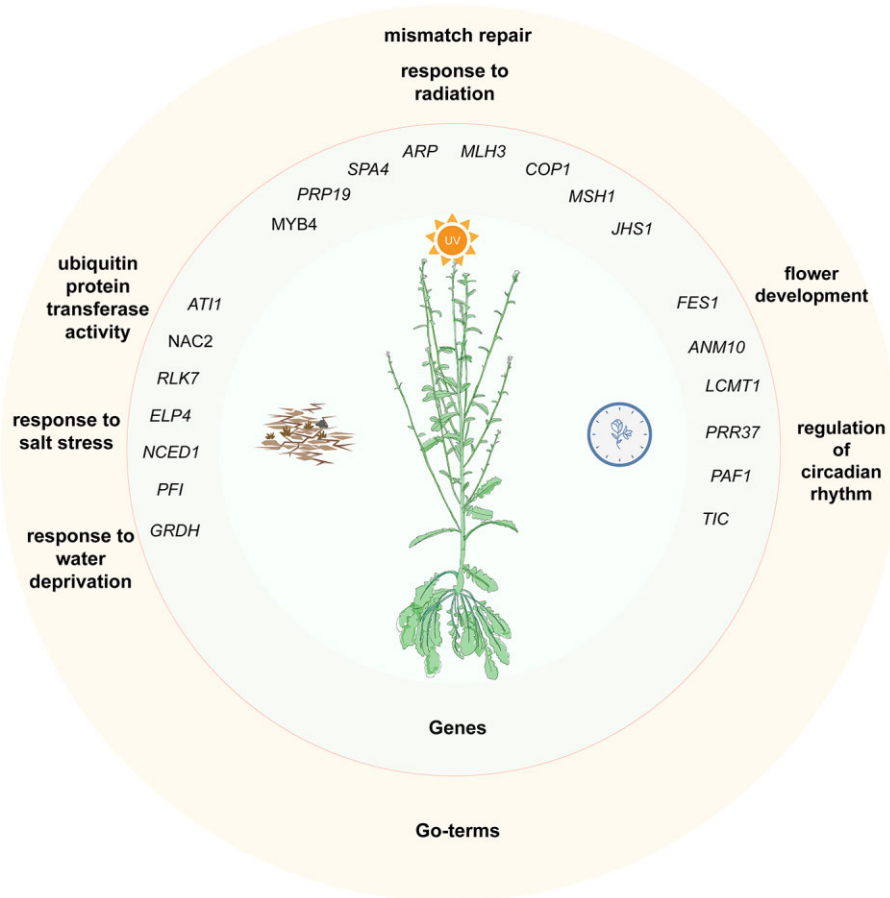
### 3.6. Self-compatibility and S-locus structure of *Crucihimalaya lasiocarpa*

Self-pollination and self-compatibility (SC) in Brassicaceae arise through functional loss or disappearance of the self-incompatibility (SI) genes, including the stigma-expressed receptor kinase (*SRK*) gene at the S locus and the other pollen-expressed cysteine-rich (*SCR/SP11*) gene.<sup>73–76</sup> SI species may be vulnerable to pollen limitation due to the lack of pollinators under typical high-altitude and harsh conditions. However, species can achieve optimal reproductive fitness through self-pollination.<sup>77</sup> Using S-locus sequences from the close relatives to search against the *C. lasiocarpa* genome, we identified the *SCR* gene in our species. The *SCR* gene of *C. himalaica* had two mutations across the eight critical conserved cysteines.<sup>14</sup> We also identified these two mutations in *C. lasiocarpa*, which is critical to maintain structural and functional integrity of the *SCR* protein<sup>73–76</sup> (Supplementary Fig. S10). Using the same way, we also identified the likely *SRK* gene in *C. lasiocarpa*. Both *SRK* and *ARK3* belong to the plant receptor-like/Pelle kinase (RLKs) family.<sup>78</sup> We used the *SRK* and *ARK3* sequences from *Arabidopsis* and other species to search for the corresponding homologous sequences. We constructed the maximum likelihood tree using all homologous sequences from two *Crucihimalaya* species and other species. We found that all identified homologous sequences from *Crucihimalaya* were clustered with the *ARK3* genes of the other species and none was clustered with the *SRK* genes (Supplementary Fig. S11).

## 4. Discussion

Here, we describe a chromosomal-scale genome for an alpine plant *C. lasiocarpa* in the QTP obtained by integrating data from ONT, Hi-C and Illumina platforms. The genome assembly of *C. lasiocarpa* exhibited a total size of 255.8 Mb, and it is the first chromosomal-scale genome in the genus *Crucihimalaya*. All of our statistics indicates that the genome assembly we generated was of high completeness and continuity, ensuring the reliability of our subsequent comparative genomic analyses. Compared with previously published genome *C. himalaica* with contig N50 of 136 kb and scaffold N50 of 2 Mb, the present *C. lasiocarpa* genome has highly improved contig N50 (14 Mb) and scaffold N50 (31 Mb). Our scaffolding with Hi-C further facilitated the accurate assignment of all scaffolds to chromosomal positions. The ONT-based assembly also annotated more complete repetitive sequences than the Illumina-based assembly.<sup>79</sup> Therefore, the centromere region of the *C. lasiocarpa* genome seem to be more accurate and complete than that of *C. himalaica* (Supplementary Fig. S12). We identified substantially more TEs (especially intact LTR-RTs) in *C. lasiocarpa* than *C. himalaica*. The rapid proliferation of these repetitive elements and especially those inserted around genic regions in *C. lasiocarpa* may play a key role in promoting its genome evolution and species divergence from the congener *C. himalaica*.<sup>14</sup>

The karyotype of the genus *Capsella* is hypothesized to be similar to the inferred ancestral ACK karyotype ( $n=8$ ).<sup>66</sup> Compared with this ACK karyotype, the karyotype of our studied species *C. lasiocarpa* ( $n=8$ ) is relatively conserved because five chromosomes (Cl2, 3, 4, 5 and 7) remain stable while the other three show chromosomal structural variations, including two reciprocal translocation, two chromosome fusions and three inversions (Fig. 2d and Supplementary Fig. S9), which may have occurred since the divergence of *Crucihimalaya* from *Capsella* ~11 million years ago (MYA).<sup>80–82</sup> Another congener of the genus *Crucihimalaya*, *C. walliichii*, was also found to have a similar pericentric inversion as in *C. lasiocarpa* for the first chromosome (Supplementary Fig. S13).<sup>82</sup> These chromosomal variations may be common for various genus or



**Figure 4.** Functional adaptation of the positively selected genes (PSGs) of the genus *Crucihimalaya* to the high-altitude habitats. The outer cream-coloured circle shows examples of enriched biological process GO-terms. The inner light grey circle shows examples of candidate positive selected genes.

specific to each genus or species, since another inversion of Cl8 in *C. lasiocarpa* (Supplementary Fig. S13) is very similar to the V genomic block in the eighth chromosome of *Transberingia bursifolia* (Brassicaceae).<sup>82</sup> This ‘rare genomic changes’<sup>83</sup> may suggest phylogenetic relatedness and common ancestry, although the phylogenetic relationship between *Crucihimalaya* and *Transberingia* is unclear. Therefore, parsimonious karyotype evolution of this species (or the genus *Crucihimalaya*) may involve both fusion and fission events during reciprocal translocation and inversions (Fig. 2D). The fragments of AB genomic block in AK1 and V genomic block in AK8 broke off from the corresponding chromosomes and swapped places, and further formed chromosomes CL1 and CL8 in *C. lasiocarpa*. Future studies may still need to explore whether such a similar chromosomal rearrangement is derived from paralleling evolution or because they share the same common ancestor.

Self-pollination and SC through functional loss of the SI genes remains a common reproductive assurance for alpine plants when pollinators become scarce.<sup>77</sup> The genus *Crucihimalaya* are self-compatible while most genera of the Brassicaceae are outcrossing with strong SI.<sup>84,85</sup> In this family, SI was certified to be determined by sequence variation at a highly polymorphic S locus with *SRK* and *SCR/SP11* genes.<sup>78</sup> The *SRK* protein performs as the receptor for *SCR* to distinguish the ‘non-self’ pollen during SI response. Self-compatibility is achieved usually when *SRK* receptors could not

detect *SCR* ligands. Our study of *C. lasiocarpa* and the previous investigation of *C. himalaica*<sup>14</sup> suggested that *SRK* genes seemed to have been lost for the genus *Crucihimalaya* and the *SCR* gene also showed the relaxed selection with two mutations that disrupt the normal function of this gene. Therefore, these two evolutionary events may together lead to the self-pollination shift of the genus. In addition, the identified positively selected genes (PSGs) for the genus are involved in DNA repair, coldness- and drought-response and reproductive processes.<sup>14</sup> We also found the significantly expanded gene families involved in adaptation of alpine habitats. All of these results together suggest that the ancestor of *Crucihimalaya* might have developed such special genomic characters to adapt to the high-altitude QTP before they diversified into these two species. It remains unclear whether these two species diverged with chromosomal structural variations and what type of genes played a key role during their divergence. In order to address these questions, the chromosomal-scale genome sequence of *C. himalaica* needs to be assembled and population genomic data of these two species are expected to obtain in the future. In summary, the chromosomal-scale genome sequence of *C. lasiocarpa* was reported here for the first time. We also clarified the karyotype evolution for the genus of *Crucihimalaya*. Combined with the previously reported genome sequence for *C. himalaica*, we found that many genomic changes related to alpine adaptation might have occurred before the divergence of the two species.



## Acknowledgements

This study was supported by the National Natural Science Foundation of China (41771055, 31700323), National Key Research and Development programme (2017YFC0505203).

## Accession number

PRJNA763756.

## Authors' contributions

Q.H. and L.F. designed the research. Q.H. and L.F. collected the materials and performed the genome sequencing and assembly. L.F., W.Y., M.K., H.L., Z.X., T.L., X.Y., Y.R. and S.W. performed the genome annotation and evolution analyses. L.F., Q.H. and M.K. wrote the manuscript. All authors contributed to the article and approved the submitted version.

## Conflict of interest

None declared.

## Data availability

Raw sequencing data and genome assembly have been deposited at the NCBI under the BioProject PRJNA763756.

## Supplementary data

[Supplementary data](#) are available at *DNARES* online.

## References

- Norsang, G., Kocbach, L., Stamnes, J., Tsoja, W. and Pincuo, N. 2011, Spatial distribution and temporal variation of solar UV radiation over the Tibetan Plateau, *Appl. Phys. Res.*, **3**, 37.
- Mao, K., Wang, Y. and Liu, J. 2021, Evolutionary origin of species diversity on the Qinghai-Tibet Plateau, *J. Syst. Evol.*, **59**, 1142–58.
- Wang, X., Liu, S., Zuo, H., et al. 2021, Genomic basis of high-altitude adaptation in Tibetan Prunus fruit trees, *Curr. Biol.*, **31**, 3848–60.
- Guo, X., Hu, Q., Hao, G., et al. 2018, The genomes of two *Eutrema* species provide insight into plant adaptation to high altitudes, *DNA Res.*, **25**, 307–15.
- Ma, Y., Wang, J., Hu, Q., et al. 2019, Ancient introgression drives adaptation to cooler and drier mountain habitats in a cypress species complex, *Commun. Biol.*, **2**, 1–12.
- Chen, J., Huang, Y., Brachi, B., et al. 2019, Genome-wide analysis of Cushion willow provides insights into alpine plant divergence in a biodiversity hotspot, *Nat. Commun.*, **10**, 1–12.
- Li, J., Zhong, L., Wang, J., Ma, T., Mao, K. and Zhang, L. 2021, Genomic insights into speciation history and local adaptation of an alpine aspen in the Qinghai-Tibet Plateau and adjacent highlands, *J. Syst. Evol.*, **59**, 1220–31.
- Zeng, X., Yuan, H., Dong, X., et al. 2020, Genome-wide dissection of co-selected UV-B responsive pathways in the UV-B adaptation of Qingke, *Mol. Plant.*, **13**, 112–27.
- Yang, Q., Bi, H., Yang, W., et al. 2020, The genome sequence of alpine *Megacarpaea delavayi* identifies species-specific whole-genome duplication, *Front. Genet.*, **11**, 1–9.
- Li, Y., Cao, K., Li, N., et al. 2021, Genomic analyses provide insights into peach local adaptation and responses to climate change, *Genome Res.*, **31**, 592–606.
- Al-Shehbaz, I., O'Kane, S. and Price, R. 1999, Generic placement of species excluded from *Arabidopsis* (Brassicaceae), *Novon*, **9**, 296–307.
- German, D.A. and Al-Shehbaz, I.A. 2010, Nomenclatural novelties in miscellaneous Asian Brassicaceae (Cruciferae), *Nord. J. Bot.*, **28**, 646–51.
- Zhao, B., Liu, L., Tan, D. and Wang, J. 2010, Analysis of phylogenetic relationships of Brassicaceae species based on Chs sequences, *Biochem. Syst. Ecol.*, **38**, 731–9.
- Zhang, T., Qiao, Q., Novikova, P.Y., et al. 2019, Genome of *Crucihimalaya himalaica*, a close relative of *Arabidopsis*, shows ecological adaptation to high altitude, *Proc. Natl. Acad. Sci. USA*, **116**, 7137–46.
- Schranz, M.E., Lysak, M.A. and Mitchell-Olds, T. 2006, The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes, *Trends Plant Sci.*, **11**, 535–42.
- Lysak, M.A., Mandáková, T. and Schranz, M.E. 2016, Comparative paleogenomics of crucifers: ancestral genomic blocks revisited, *Curr. Opin. Plant Biol.*, **30**, 108–15.
- Kang, M., Wu, H., Yang, Q., et al. 2020, A chromosome-scale genome assembly of *Isatis indigotica*, an important medicinal plant used in traditional Chinese medicine: an *Isatis* genome, *Hortic. Res.*, **7**, 1–10.
- Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y. and Dekker, J. 2012, Hi-C: a comprehensive technique to capture the conformation of genomes, *Methods*, **58**, 268–76.
- Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., et al. 2009, Comprehensive mapping of long-range interactions reveals folding principles of the human genome, *Science* (80), **326**, 289–93.
- Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. 2002, Capturing chromosome conformation, *Science* (80), **295**, 1306–11.
- Hu, J., Fan, J., Sun, Z. and Liu, S. 2020, NextPolish: a fast and efficient genome polishing tool for long-read assembly, *Bioinformatics*, **36**, 2253–5.
- Durand, N.C., Shamim, M.S., Machol, I., et al. 2016, Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments, *Cell Syst.*, **3**, 95–8.
- Durand, N.C., Robinson, J.T., Shamim, M.S., et al. 2016, Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom, *Cell Syst.*, **3**, 99–101.
- Dudchenko, O., Batra, S.S., Omer, A.D., et al. 2017, De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds, *Science*, **356**, 92–5.
- Dudchenko, O., Shamim, M.S., Batra, S.S., et al. 2018, The juicebox assembly tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under \$1000. *bioRxiv*. 10.1101/254797
- Li, H. 2013, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, *arXiv*, 1303.3997v2..
- Parra, G., Bradnam, K. and Korf, I. 2007, CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes, *Bioinformatics*, **23**, 1061–7.
- Ou, S. and Jiang, N. 2018, LTR\_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons, *Plant Physiol.*, **176**, 1410–22.
- Price, A.L., Jones, N.C. and Pevzner, P.A. 2005, De novo identification of repeat families in large genomes, *Bioinformatics*, **21**, i351–8.
- Chen Nansheng, M.T. 2009, Using RepeatMasker to identify repetitive elements in genomic sequences, *Curr. Protoc. Bioinforma.*, **5**, 1–14.
- Ellinghaus, D., Kurtz, S. and Willhoeft, U. 2008, LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons, *BMC Bioinformatics*, **9**, 1–14.
- Xu, Z. and Wang, H. 2007, LTR-FINDER: an efficient tool for the prediction of full-length LTR retrotransposons, *Nucleic Acids Res.*, **35**, 265–8.
- Ma, B., Kuang, L., Xin, Y. and He, N. 2019, New insights into long terminal repeat retrotransposons in mulberry species, *Genes (Basel)*, **10**, 285.
- Bolger, A.M., Lohse, M. and Usadel, B. 2014, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, **30**, 2114–20.

35. Haas, B.J., Papanicolaou, A., Yassour, M., et al. 2013, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, *Nat. Protoc.*, **8**, 1494–512.
36. Haas, B.J., Delcher, A.L., Mount, S.M., et al. 2003, Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies, *Nucleic Acids Res.*, **31**, 5654–66.
37. Slater, G.S.C. and Birney, E. 2005, Automated generation of heuristics for biological sequence comparison, *BMC Bioinformatics*, **6**, 31–11.
38. Stanke, M., Steinkamp, R., Waack, S. and Morgenstern, B. 2004, AUGUSTUS: a web server for gene finding in eukaryotes, *Nucleic Acids Res.*, **32**, 309–12.
39. Haas, B.J., Salzberg, S.L., Zhu, W., et al. 2008, Automated eukaryotic gene structure annotation using EVIDENCEModeler and the program to assemble spliced alignments, *Genome Biol.*, **9**, R7–22.
40. Zdobnov, E.M. and Apweiler, R. 2001, InterProScan - an integration platform for the signature-recognition methods in InterPro, *Bioinformatics*, **17**, 847–8.
41. Bairoch, A. and Apweiler, R. 2000, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res.*, **28**, 45–8.
42. Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M. 2005, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics*, **21**, 3674–6.
43. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. 2012, KEGG for integration and interpretation of large-scale molecular data sets, *Nucleic Acids Res.*, **40**, 109–14.
44. Mandáková, T., Zozomová-Lihová, J., Kudoh, H., Zhao, Y., Lysak, M.A. and Marhold, K. 2019, The story of promiscuous crucifers: origin and genome evolution of an invasive species, *Cardamine occulta* (Brassicaceae), and its relatives, *Ann. Bot.*, **124**, 209–20.
45. Wang, Y., Tang, H., Debarry, J.D., et al. 2012, MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity, *Nucleic Acids Res.*, **40**, 1–14.
46. Kielbasa, S.M., Wan, R., Sato, K., Horton, P. and Frith, M.C. 2011, Adaptive seeds tame genomic sequence comparison, *Genome Res.*, **21**, 487–93.
47. Sun, P., Jiao, B., Yang, Y., et al. 2021, WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *bioRxiv* 10.1101/2021.04.29.441969.
48. Emms, D.M. and Kelly, S. 2019, OrthoFinder: phylogenetic orthology inference for comparative genomics, *Genome Biol.*, **20**, 1–14.
49. Katoh, K. and Standley, D.M. 2013, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Mol. Biol. Evol.*, **30**, 772–80.
50. Castresana, J. 2000, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, *Mol. Biol. Evol.*, **17**, 540–52.
51. Stamatakis, A. 2014, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, *Bioinformatics*, **30**, 1312–3.
52. Yang, Z. 2007, PAML 4: phylogenetic analysis by maximum likelihood, *Mol. Biol. Evol.*, **24**, 1586–91.
53. Yang, Z. and Rannala, B. 1997, Monte Carlo method A Markov Chain I-I, *Integr. VLSI J.*, **14**, 717–24.
54. Hedges, S.B., Dudley, J. and Kumar, S. 2006, TimeTree: a public knowledge-base of divergence times among organisms, *Bioinformatics*, **22**, 2971–2.
55. De Bie, T., Cristianini, N., Demuth, J.P. and Hahn, M.W. 2006, CAFE: a computational tool for the study of gene family evolution, *Bioinformatics*, **22**, 1269–71.
56. Xie, C., Mao, X., Huang, J., et al. 2011, KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases, *Nucleic Acids Res.*, **39**, W316–22.
57. Yang, Z., Wong, W.S.W. and Nielsen, R. 2005, Bayes empirical Bayes inference of amino acid sites under positive selection, *Mol. Biol. Evol.*, **22**, 1107–18.
58. Woodard, S.H., Fischman, B.J., Venkat, A., et al. 2011, Genes involved in convergent evolution of eusociality in bees, *Proc. Natl. Acad. Sci. USA*, **108**, 7472–7.
59. Wang, H.Y., Tang, H., Shen, C.K.J. and Wu, C.I. 2003, Rapidly evolving genes in human. I. The glycoporphins and their possible role in evading malaria parasites, *Mol. Biol. Evol.*, **20**, 1795–804.
60. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, **31**, 3210–2.
61. Ou, S., Chen, J. and Jiang, N. 2018, Assessing genome assembly quality using the LTR Assembly Index (LAI), *Nucleic Acids Res.*, **46**, e126.
62. Maumus, F. and Quesneville, H. 2014, Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana, *Nat. Commun.*, **5**, 1–9.
63. Slotte, T., Hazzouri, K.M., Ågren, J.A., et al. 2013, The Capsella rubella genome and the genomic consequences of rapid mating system evolution, *Nat. Genet.*, **45**, 831–5.
64. Lee, H., Ayarpadikannan, S. and Kim, H. 2015, Role of transposable elements in genomic rearrangement, evolution, gene regulation and epigenetics in primates, *Genes Genet. Syst.*, **90**, 245–57.
65. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. and Frazer, K.A. 2006, Common deletions and SNPs are in linkage disequilibrium in the human genome, *Nat. Genet.*, **38**, 82–5.
66. Lysak, M.A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K. and Schubert, I. 2006, Mechanisms of chromosome number reduction in Arabidopsis thaliana and related Brassicaceae species, *Proc. Natl. Acad. Sci. USA*, **103**, 5224–9.
67. Lv, S., Cheng, S., Wang, Z., et al. 2020, Draft genome of the famous ornamental plant *Paeonia suffruticosa*, *Ecol. Evol.*, **10**, 4518–30.
68. Long, M., Betrán, E., Thornton, K. and Wang, W. 2003, The origin of new genes: glimpses from the young and old, *Nat. Rev. Genet.*, **4**, 865–75.
69. Conant, G.C. and Wolfe, K.H. 2007, Increased glycolytic flux as an outcome of whole-genome duplication in yeast, *Mol. Syst. Biol.*, **3**, 129.
70. Tissot, N. and Ulm, R. 2020, Cryptochrome-mediated blue-light signaling modulates UVR8 photoreceptor activity and contributes to UV-B tolerance in Arabidopsis, *Nat. Commun.*, **11**, 1–10.
71. Zhou, J., Wang, X., He, Y., et al. 2020, Differential phosphorylation of the transcription factor WRKY33 by the protein kinases CPK5/CPK6 and MPK3/MPK6 cooperatively regulates camalexin biosynthesis in Arabidopsis, *Plant Cell*, **32**, 2621–38.
72. Schuëgger, R., Nafisi, M., Mansourova, M., et al. 2006, CYP71B15 (PAD3) catalyzes the final step in camalexin biosynthesis, *Plant Physiol.*, **141**, 1248–54.
73. Lin, J., Wang, D., Chen, X., et al. 2017, An (E,E)- $\alpha$ -farnesene synthase gene of soybean has a role in defence against nematodes and is involved in synthesizing insect-induced volatiles, *Plant Biotechnol. J.*, **15**, 510–9.
74. Li, W., Li, K., Huang, Y., et al. 2020, SMRT sequencing of the *Oryza rufipogon* genome reveals the genomic basis of rice adaptation, *Commun. Biol.*, **3**, 1–11.
75. Moutinho, A.F., Bataillon, T. and Duthel, J.Y. 2020, Variation of the adaptive substitution rate between species and within genomes, *Ecol. Evol.*, **34**, 315–38.
76. Zhu, J., Jae, C.J., Zhu, Y., et al. 2008, Involvement of Arabidopsis HOS15 in histone deacetylation and cold tolerance, *Proc. Natl. Acad. Sci. USA*, **105**, 4945–50.
77. Zhao, Z. and Wang, Y. 2015, Selection by pollinators on floral traits in generalized *Trollius ranunculoides* (Ranunculaceae) along altitudinal gradients, *PLoS One*, **10**, e0118299.
78. Kitashiba, H. and Nasrallah, J.B. 2014, Self-incompatibility in Brassicaceae crops: lessons for interspecific incompatibility, *Breed. Sci.*, **64**, 23–37.
79. Roscito, J.G., Sameith, K., Pippel, M., et al. 2018, The genome of the tegu lizard *Salvator merianae*: combining Illumina, PacBio, and optical mapping data to generate a highly contiguous assembly, *Gigascience*, **7**, 1–13.
80. Hohmann, N., Wolf, E.M., Lysak, M.A. and Koch, M.A. 2015, A time-calibrated road map of Brassicaceae species radiation and evolutionary history, *Plant Cell*, **27**, 2770–84.
81. Huang, C.H., Sun, R., Hu, Y., et al. 2016, Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and

- 
- supports convergent morphological evolution, *Mol. Biol. Evol.*, **33**, 394–412.
82. Mandáková, T., Joly, S., Krzywinski, M., Mummenhoff, K. and Lysak, M.A. 2010, Fast diploidization in close mesopolyploid relatives of *Arabidopsis*, *Plant Cell.*, **22**, 2277–90.
83. Rokas, A. and Holland, P.W.H. 2000, Rare genomic changes as a tool for phylogenetics, *Trends Ecol. Evol.*, **15**, 454–9.
84. Roy, S., Ueda, M., Kadowaki, K. and Tsutsumi, N. 2010, Different status of the gene for ribosomal protein S16 in the chloroplast genome during evolution of the genus *Arabidopsis* and closely related species, *Genes Genet. Syst.*, **85**, 319–26.
85. Tedder, A., Ansell, S.W., Lao, X., Vogel, J.C. and Mable, B.K. 2011, Sporophytic self-incompatibility genes and mating system variation in *Arabis alpina*, *Ann. Bot.*, **108**, 699–713.