Genome of *Crucihimalaya himalaica*, a close relative of *Arabidopsis*, shows ecological adaptation to high altitude

Ticao Zhang,a,b,1 Qin Qiao,a,1 Polina Yu. Novikova,c,e,1 Qia Wang,a Jipei Yue,a Yanlong Guan,b Shengping Ming,b Tianmeng Liu,a Ji De,b Yixuan Liu,b Ihsan A. Al-Shehbaz,b Hang Sun,b Marc Van Montagu,c,e,2 Jinling Huang,a,h,i,2 Yves Van de Peer,d,e,j,2 and La Qiangd,2

*Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, 650201 Kunming, China; College of Chinese Material Medica, Yunnan University of Chinese Medicine, 650050 Kunming, China; School of Agriculture, Yunnan University, 650091 Kunming, China; Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium; *Center for Plant Systems Biology, VIB, 9052 Ghent, Belgium; *Institute of Biodiversity Science and Geobiology, College of Science, Tibet University, 850012 Lhasa, China; †Missouri Botanical Garden, St. Louis, MO 63166; ‡Department of Biology, East Carolina University, Greenville, NC 27858; †Institute of Plant Stress Biology, State Key Laboratory of Cotton Biology, School of Life Sciences, Henan University, 475001 Kaifeng, China; and §Department of Biochemistry, Genetics and Microbiology, University of Pretoria, 0028 Pretoria, South Africa

Contributed by Marc Van Montagu, February 21, 2019 (sent for review October 19, 2018; reviewed by Ya-Long Guo, Matthew B. Hufferd, and Martin Lascoux)

**Crucihimalaya himalaica**, a close relative of *Arabidopsis* and *Capsella*, grows on the Qinghai-Tibet Plateau (QTP) about 4,000 m above sea level and represents an attractive model system for studying speciation and ecological adaptation in extreme environments. We assembled a draft genome sequence of 234.72 Mb encoding 27,019 genes and investigated its origin and adaptive evolutionary mechanisms. Phylogenomic analyses based on 4,586 single-copy genes revealed that *C. himalaica* is most closely related to *Capsella* (estimated divergence 8.8 to 12.2 Mya), whereas both species form a sister clade to *Arabidopsis thaliana* and *Arabidopsis lyrata*, from which they diverged between 12.7 and 17.2 Mya. LTR retrotransposons in *C. himalaica* proliferated shortly after the dramatic uplift and climatic change of the Himalayas from the Late Pliocene to Pleistocene. Compared with closely related species, *C. himalaica* showed significant contraction and pseudogenization in gene families associated with disease resistance and also significant expansion in gene families associated with ubiquitin-mediated proteolysis and DNA repair. We identified hundreds of genes involved in DNA repair, ubiquitin-mediated proteolysis, and reproduct processes with signs of positive selection. Gene families showing dramatic changes in size and genes showing signs of positive selection are likely candidates for *C. himalaica*’s adaptation to intense radiation, low temperature, and pathogen-depauperate environments in the QTP. Loss of function at the S-locus, the reason for the transition to self-fertilization of *C. himalaica*, might have enabled its QTP occupation. Overall, the genome sequence of *C. himalaica* provides insights into the mechanisms of plant adaptation to extreme environments.

Significance

**Crucihimalaya himalaica** is a close relative of *Arabidopsis* with typical Qinghai-Tibet Plateau (QTP) distribution. Here, by combining short- and long-read sequencing technologies, we provide a new genome sequence of *C. himalaica*. Our results suggest that the quick uplifting of the QTP coincided with the expansion of repeat elements. Gene families showing dramatic contractions and expansions, as well as genes showing clear signs of natural selection, were likely responsible for *C. himalaica*’s specific adaptation to the harsh environment of the QTP. We also show that the transition to self-pollination of *C. himalaica* might have enabled its occupation of the QTP. This study provides insights into how plants might adapt to extreme environmental conditions.


Reviewer: Y.-L.G. Chinese Academy of Sciences; M.B.H., Iowa State University; and M.L., Uppsala University.

The authors declare no conflict of interest.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: The raw genomic reads generated in this study have been deposited in the NCBI Sequence Read Archive (BioProject PRJNA521295).

1T.Z., Q.Q., and P.Y.N. contributed equally to this work.

2To whom correspondence may be addressed. Email: marc.vanmontagu@vib-ugent.be, huang@ecu.edu, yves.vandepeer@vib-ugent.be, or lqiangchong@163.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1817580116/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1817580116

PNAS Latest Articles | 1 of 10
C. himalaica originated about 3.56 Mya (12). Given the postulated timing of the most recent rapid uplift of the QTP during the period of late Miocene to late Pliocene (1, 13), many of the genetic changes in C. himalaica might reflect its adaptation to the extreme environment of QTP (4).

In a previous study (4), we identified 487 positively selected genes (PSGs) in the transcriptome of C. himalaica. Predicted functions of these PSGs indicate that they potentially contribute to miscellaneous traits of adaptive importance, such as response to UV radiation, DNA repair, and membrane stabilization, which presumably are important for the adaptation of C. himalaica to the specialized environment on the QTP. Therefore, we believe that this species represents a model system for the study of speciation and ecological adaptation in extreme environments. In the present study, we generated de novo whole-genome sequences of C. himalaica and applied comparative and evolutionary genomics approaches to clarify the origin of this species and to investigate signals of adaptation. Our goal is to gain a deeper understanding into the mechanisms by which C. himalaica has adapted to the complex extreme conditions on the QTP at the whole-genome level.

Results and Discussion

Genome Assembly and Annotation. The genome size of C. himalaica was estimated to be 265.23 Mb with a heterozygosity of 0.70% on the basis of k-mer statistics (14) (SI Appendix, Table S1). In total, we generated 50.68 Gb of paired-end (PE) and mate-pair (MP) reads with different insert sizes using an Illumina HiSeq platform combined with PacBio sequencing technology (SI Appendix, Table S2). A total assembly of 234.72 Mb, consisting of 583 scaffolds (scaffold N50 length, 2.09 Mb; longest scaffold, 8.34 Mb), was achieved by combination of the ALLPATHS-LG (15) and PB Jelly2 (16) assembly strategies (Table 1). To assess assembly accuracy, we remapped raw sequencing reads of a small fragment library to the assembled genome. With a 96.69% mapping rate and 224.57× average sequence depth, the reads covered 96.43% of the genome, which implies that the current assembly covered almost all unique genomic regions. About 94.05% of the assembly was covered by at least 20× reads, which guaranteed a highly accurate assembly at the single-nucleotide level (SI Appendix, Table S3).

A total of 27,019 protein-coding genes were predicted (Table 1 and SI Appendix, Table S4). Among these, 26,806 genes (99.21%) were functionally annotated (SI Appendix, Table S5). In addition to protein-coding genes, various noncoding RNA sequences were identified and annotated (SI Appendix, Table S6), including 448 microRNAs, 577 transfer RNAs, 153 ribosomal RNAs, and 974 small nuclear RNAs. Gene region completeness was evaluated by RNA sequencing data (SI Appendix, Table S7). Of the 29,420 transcripts assembled by Trinity, 99.74% were mapped to the genome assembly, and 92.37% were complete. The completeness of gene regions was further assessed using BUSCO (Benchmarking Universal Single Copy Orthologs) (17), which showed that 96% of the plant single-copy orthologs were complete (SI Appendix, Table S8). Compared with other close relatives, the C. himalaica genome contains a similar number of transcription factors (1,711) and transcription regulators (413) (SI Appendix, Table S9). The C. himalaica genome showed strong homology with the genome of A. thaliana, Arabidopsis lyrata, and Capsella rubella, except in centromeres and certain high-repeat regions (Fig. 1 and SI Appendix, Figs. S1–S3). These results implied that the assembly was of high quality, thus ensuring the reliability of subsequent comparative genomic analyses in this study.

Repeat Elements. Compared with the genome size of closely related species, such as C. rubella (which has an assembled genome size of 134.8 Mb, whereas the genome size estimated from flow cytometry was 219 Mb) (18) and A. thaliana [125 Mb (19) and the estimated genome size of the reference accession Col-0 from flow cytometry is 166 Mb (20)], the C. himalaica genome is considerably larger. It has been well documented that polyploidization (whole-genome duplication, WGD) events and transposable element (TE) amplification are major causes of genome expansion (21, 22). Analyses of age distributions built from transversion substitutions at fourfold degenerate sites (4DTv) indicated that, except for the a (4DTv distance = ~0.3) and β (4DTv distance = ~0.6) polyomiodized events shared among members of the Brassicaeae (23), C. himalaica has not undergone an additional species-specific WGD event (Fig. 24). Next, we investigated the content and evolutionary history of TEs in C. himalaica. Using de novo prediction of TEs (Materials and Methods), we identified and marked 46.91% of the assembly as repeat regions, among which TEs occupied 45.78% of the genome assembly length (SI Appendix, Table S10), which is higher than that reported for A. lyrata and A. thaliana (29.7% and 23.7%, respectively) (24). This was also apparent in a CIRCOS genomic comparison plot, which showed that the density of TEs in C. himalaica was higher than that in A. thaliana or C. rubella, whereas the density of genes in C. himalaica was lower than that in A. thaliana or C. rubella (Fig. 1).

In particular, a high proportion of TEs in C. himalaica were LTR retrotransposons (30.37%), whereas other retrotransposons (short and long interspersed nuclear elements) only constituted 3.31% collectively (SI Appendix, Table S10). The retrotransposon proliferation might be responsible for the genome-size expansion in C. himalaica. To investigate the evolutionary dynamics of LTR retrotransposons, we estimated their insertion dates in four closely related species (Materials and Methods). Compared with A. thaliana, which shows a large number of microdeletions in noncoding DNA and transposons (24), A. lyrata has a comparatively high proportion of recent insertions (18) (Fig. 2B), possibly contributing to its larger genome size (207 Mb). The proliferation of LTR retrotransposons in C. himalaica (Fig. 2B and SI Appendix, Fig. S4 and Table S11) peaks at ~2.0 Mya, shortly after the dramatic elevational and climatic changes of the QTP between the late Miocene and late Pliocene (1, 13). The activity of TEs (including LTR retrotransposons) can lead to diverse genetic changes (e.g., chromosomal rearrangements and gene duplication, creation, and disruption) that may drive lineage-specific diversification and adaptation (25, 26). We therefore hypothesize that the proliferation of LTR retrotransposons likely contributed to the diversification, speciation, and adaptation of C. himalaica, similar to the proliferation of BARE-1 retrotransposons in wild barley (27) and DIRS1 retrotransposons in Antarctic teleost (28), which might have facilitated the adaptation of plants to drier slopes and subzero temperatures, respectively. Another example from apples involves a major burst of retrotransposons ~21.0 Mya, which coincided with the uplift of the Tian Shan mountains, the postulated center of origin of apples (29).

Phylogenetic Tree Construction and Estimation of Divergence Times. Previous studies suggest that Crataeghimala is an endemic genus in the Himalayas and is most closely related to Arabidopsis

<table>
<thead>
<tr>
<th>Genome features</th>
<th>Contigs</th>
<th>Scaffolds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total length, bp</td>
<td>230,905,116</td>
<td>234,722,603</td>
</tr>
<tr>
<td>Total no.</td>
<td>3,983</td>
<td>583</td>
</tr>
<tr>
<td>Longest length, bp</td>
<td>1,756,581</td>
<td>8,343,586</td>
</tr>
<tr>
<td>No. of length ≥2,000</td>
<td>3,586</td>
<td>429</td>
</tr>
<tr>
<td>Length of N50, bp</td>
<td>136,392</td>
<td>2,088,603</td>
</tr>
<tr>
<td>No. of N50</td>
<td>406</td>
<td>34</td>
</tr>
<tr>
<td>Length of N90, bp</td>
<td>32,421</td>
<td>470,087</td>
</tr>
<tr>
<td>No. of N90</td>
<td>1,711</td>
<td>129</td>
</tr>
<tr>
<td>GC content, %</td>
<td>36.38</td>
<td>—</td>
</tr>
<tr>
<td>No. of genes</td>
<td>—</td>
<td>27,019</td>
</tr>
</tbody>
</table>
Fig. 1. Comparative analyses of genomic features of *C. himalaica* vs. *A. thaliana* (A) and *C. himalaica* vs. *C. rubella* (B). Tracks from inside to outside are collinearity between both genomes, number of chromosomes/scaffolds, gene density, and TE density.

(30, 31). However, our earlier phylogenetic analyses of transcriptome sequences found that *C. himalaica* formed a clade sister to *C. rubella* (4). Applying OrthoMCL (32) to nine published whole-genome sequences from Brassicaceae (*SI Appendix, Table S12*), we identified a total of 22,670 orthogroups. Among these orthogroups, 4,586 contained putative single-copy gene families. To verify the phylogenetic position of *C. himalaica*, we generated a maximum likelihood phylogenetic tree with a trimmed and concatenated protein sequence alignment from 4,586 single-copy genes in nine species. The resulting phylogeny indicated that *C. himalaica* was most closely related to *C. rubella*, and that these two species in turn formed a clade with *A. lyrata* and *A. thaliana* (Fig. 2C), confirming our previous results (4). The above-mentioned four genera (which are classified in the tribe Camelinae), together with the allied *Cardamine hirsuta*, were often recognized as Lineage I or Clade A in previous phylogenetic studies (33, 34).

*C. himalaica* and *C. rubella* were estimated to have diverged ~10.6 (8.8 to 12.2) Mya in our analyses using MCMCtree (35) with three calibration points (*Materials and Methods* and Fig. 2C); the two species diverged from *Arabidopsis* ~14.6 (12.7 to 17.2) Mya and from *Brassica rapa* ~23.6 (20.9 to 27.8) Mya. These results are in agreement with previous estimates (11, 36). Given the absence of genomic information for the remaining species of *Crucihimalaya*, we are unable to date the exact origin of *C. himalaica*, which was estimated previously at ~3.56 Mya (11, 12). However, *C. himalaica* must have evolved after the split with *Capsella* and almost certainly less than 10.6 Mya. Such estimation is in accordance with the timing of the most recent rapid uplift of the QTP from late Miocene to late Pliocene (1, 13). The genus *Crucihimalaya* is assigned to the tribe Crucihimalayae, which includes two additional genera, *Ladakiella* and *Transberingia* (37, 38). The tribe Crucihimalayae is mainly distributed in Central Asia, with one species (*Transberingia bursifolia*) extending into North America. Therefore, we speculate that the ancestor of *C. himalaica* dispersed from Central Asia into the Himalayas, and speciation of *C. himalaica* was likely triggered by the rapid uplift of the Himalayas, which resulted in the evolution of specialized phenotypic and physiological characters as adaptations to the extreme environment (discussed further below).

**Gene Family Expansion and Contraction.** Significant expansion or contraction in the size of particular gene families is often associated with the adaptive divergence of closely related species (39, 40). Comparisons of the genomes of *C. himalaica* and four close relatives (Fig. 2D) identified a total of 151 gene families that are significantly (*P < 0.01*) expanded in *C. himalaica* and 89 gene families that are significantly contracted (*SI Appendix, Table S13*). Based on Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) annotations, expanded gene families were highly enriched in ubiquitin-conjugating enzyme (E2) activity (*P = 7.17E-40*) and DNA repair pathways (*P = 1.45E-12*) (Table 2). It is notable that the most significantly contracted gene families in the *C. himalaica* genome were found to be functionally related to disease and immune responses, such as the Toll-like receptor and NF-κB signaling pathway (Table 2). The majority of bacteria interact with Toll-like receptors on the surface of the host cell membrane, stimulating the NF-κB signaling pathway in the immune response (41). The Toll and interleukin-1 receptor (TIR) is an N-terminal component of the nucleotide-binding site (NBS) disease resistance protein family, which includes the TIR-NBS-LRR (TNL) and CC-NBS-LRR (CNL) subfamilies. In the *C. himalaica* genome, both TNL and CNL subfamilies underwent severe contractions compared with their close relatives (Fig. 3D). In particular, disease resistance RPP8-like proteins (belonging to CNL) have undergone the most significant contraction in the *C. himalaica* genome (6 vs. 16–29 members, *P = 1.26E-10*). These proteins are major players in plant defense against pathogens by triggering hypersensitive responses (42). The rapid evolutionary expansion or contraction of the NBS gene family may be a fundamental strategy for plants to adapt to the rapidly changing species-specific pathogen spectrum (43, 44). As fewer microorganisms exist on the QTP owing to the harsh environments characterized by cold temperatures, aridity, and high UV radiation (45), it is reasonable to presume that the reduction in number of NBS genes in the *C. himalaica* genome is due to a lighter load of...
pathogens present in the environment and therefore the scarcity of pathogen infection on the QTP. We performed additional searches for NBS pseudogenes, which were assumed to have frameshift mutations and/or premature stop codons. Only one pseudogene could be found in *A. thaliana* and *A. lyrata*, but 20, 15, and 7 NBS pseudogenes could be identified in the *C. himalaica*, *C. rubella*, and *C. hirsuta* genomes, respectively (Fig. 3B). Therefore, pseudogenization of NBS disease resistance genes in *C. himalaica* is at least partly responsible for the contraction of this gene family. Similar observations were made for the genome of the ground tit (*Parus humilis*), native to the QTP, where MHC genes involved in the cellular immune defense against pathogens also show significant contractions (46). Although the mechanism of MHC gene evolution in ground tit remains unclear, it is worth noting that both ground tit and *C. himalaica* show a convergent evolution with respect to the contraction of gene families involved in defense against pathogens.

**Positive Selection on Single-Copy Genes.** Orthologs that show signs of positive selection usually underwent adaptive divergence (47). Previously, we observed that the ratio of nonsynonymous to synonymous substitutions (dN/dS or ω) in *C. himalaica* is higher than those of closely related species, suggestive of accelerated evolution in *C. himalaica* after divergence from its ancestral lineage (4). In the present study, we conducted a positive selection analysis using the genomic sequences of *C. himalaica* and four close relatives. Among the 21,383 orthogroups, 11,085 contained single-copy orthologous genes. We used the branch-site model of the PAML 4 package (35) to identify genes with signs of positive selection. As a result, 844 genes possibly under positive selection were identified in the *C. himalaica* genome (ω > 1, P < 0.05). Of these genes, 610 showed highly significant (P < 0.01) positive selection (SI Appendix, Table S14). A KEGG functional classification of the 610 significant PSGs in the *C. himalaica* genome (SI Appendix, Table S15) showed that several categories associated with DNA repair, the ubiquitin system, as well as plant hormone biosynthesis and signal transduction were enriched. Genes involved in DNA repair were also identified as being under positive selection pressure in a previous transcriptomic study of *C. himalaica* (4). It is notable that significantly expanded gene families and PSGs were both enriched in DNA repair and protein ubiquitination pathways. Signal transduction-related CheY-like genes were also found to have undergone both significant positive selection and expansion events in our previous study on the adaptive evolution of the cyanobacterium *Trichormus* sp. NMC-1 on the QTP (3), again providing evidence that expansion/duplication and (subsequent) positive selection of genes are an important mechanism for plant adaptive evolution.

The extremely intense UV radiation on the QTP may influence plant growth and developmental processes such as photoperiodism and flowering, or cause DNA, RNA, and protein...
damage (48). Our results showed that 10 PSGs in the *C. himalaica* genome are involved in the DNA repair pathway (SI Appendix, Table S15). Notably, several genes in the nucleotide excision repair pathway showed signatures of positive selection. One of them encodes the DNA excision repair protein ERCC-1 (or UV hypersensitive 7), which was also identified previously (4) and is reported to function as a component of a structure-specific endonuclease that cleaves on the 5' side of UV photoproducts in DNA (49). In addition to ERCC-1, we identified PSGs that encode the AP endonuclease 2 protein (APEX2), DNA polymerase delta subunit 2 (POLD2), DNA mismatch repair protein MLH1, and DNA repair and recombination protein RAD54 (SI Appendix, Table S14). These PSGs are involved in base excision repair, mismatch excision repair, and homologous recombination, suggesting that *C. himalaica* has evolved an integrated DNA-repair mechanism to adapt to the harsh habitats caused by the uplift of the QTP. Moreover, high UV-B radiation is a common stress that both animals and plants on the QTP must cope with. The DNA repair and radiation responses pathways similarly have played crucial roles in the highland adaptation of the Tibetan highland barley (6), Tibetan antelope (50), Tibetan chicken (51), and Tibetan hot-spring snake (52).

Our analyses also indicated that ubiquitin system-related gene families underwent significant expansion and natural selection. Ubiquitin-mediated proteolysis impacts almost every aspect of plant growth and development, including plant hormone signal transduction, photomorphogenesis, reproduction, and abiotic stress responses (53). Such a multifunctional biological process is of even greater importance for plants in the complex harsh environment on the QTP. Among the seven PSGs in the ubiquitin system based on the KEGG annotation, for instance, one encodes an ortholog of the COP10 protein in *A. thaliana*, which can enhance the activity of ubiquitin-conjugating enzymes (E2s) (54) and participates in light signal transduction and photomorphogenesis (55). An additional PSG encodes the ortholog of the auxin transport protein BIG (E3 ubiquitin-protein ligase UBR4) in *A. thaliana*, which not only influences multiple auxin-mediated developmental processes (e.g., lateral root production and

<table>
<thead>
<tr>
<th>Gene families</th>
<th>Expanded</th>
<th>Contracted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ubiquitin-mediated proteolysis</td>
<td>19</td>
<td>64</td>
</tr>
<tr>
<td>Mismatch repair</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>Homologous recombination</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>DNA replication</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>Nucleotide excision repair</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>RNA polymerase</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>Pyrimidine metabolism</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>Purine metabolism</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>NF-κB signaling pathway</td>
<td>64</td>
<td>31</td>
</tr>
<tr>
<td>Toll-like receptor signaling pathway</td>
<td>64</td>
<td>12</td>
</tr>
<tr>
<td>Retinol metabolism</td>
<td>21</td>
<td>10</td>
</tr>
<tr>
<td>Metabolism of xenobiotics by P450</td>
<td>18</td>
<td>103</td>
</tr>
<tr>
<td>Tryptophan metabolism</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>Steroid hormone biosynthesis</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Metabolic pathways</td>
<td>31</td>
<td>1,243</td>
</tr>
<tr>
<td>Arachidonic acid metabolism</td>
<td>10</td>
<td>103</td>
</tr>
<tr>
<td>Linoleic acid metabolism</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Lysosome</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Apoptosis</td>
<td>4</td>
<td>33</td>
</tr>
<tr>
<td>Apoptosis</td>
<td>4</td>
<td>33</td>
</tr>
</tbody>
</table>

Table 2. Functional annotation of the most significantly expansive and contract gene families in *C. himalaica*

<table>
<thead>
<tr>
<th>Gene families</th>
<th>Expanded</th>
<th>Contracted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ubiquitin-mediated proteolysis</td>
<td>19</td>
<td>64</td>
</tr>
<tr>
<td>Mismatch repair</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>Homologous recombination</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>DNA replication</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>Nucleotide excision repair</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>RNA polymerase</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>Pyrimidine metabolism</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>Purine metabolism</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>NF-κB signaling pathway</td>
<td>64</td>
<td>31</td>
</tr>
<tr>
<td>Toll-like receptor signaling pathway</td>
<td>64</td>
<td>12</td>
</tr>
<tr>
<td>Retinol metabolism</td>
<td>21</td>
<td>10</td>
</tr>
<tr>
<td>Metabolism of xenobiotics by P450</td>
<td>18</td>
<td>103</td>
</tr>
<tr>
<td>Tryptophan metabolism</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>Steroid hormone biosynthesis</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Metabolic pathways</td>
<td>31</td>
<td>1,243</td>
</tr>
<tr>
<td>Arachidonic acid metabolism</td>
<td>10</td>
<td>103</td>
</tr>
<tr>
<td>Linoleic acid metabolism</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Lysosome</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Apoptosis</td>
<td>4</td>
<td>33</td>
</tr>
</tbody>
</table>

Fig. 3. Size of the NBS gene family (A) and number of pseudo-NBS genes (B) in the *C. himalaica* genome compared with those of related species.
inflorescence architecture) but also plays a critical role in a multitude of light and phytohormone pathways (56). These PSGs enriched in the ubiquitin system might be important for Crucihimalaya to better survive the harsh environment in the QTP. Similarly, in three mangrove species, the ubiquitin-mediated proteolysis pathway includes 12 genes that have experienced convergent evolution at conservative sites, presumably important for adaptation of mangroves to the harsh coastal habitat (57).

Moreover, many of the identified PSGs were associated with reproduction pathways (92 PSGs, P = 5.40E-06) based on GO annotation ([SI Appendix, Table S16]). One of the PSGs encodes the phytochrome and flowering time regulatory protein 1 (PFT1), which represses the PhyB-mediated light signaling and regulates the expression of FLOWERING LOCUS T (FT) and CONSTANS (CO) (58, 59). In a low-red to far-red ratio environment, plants require CO and FT to fully accelerate flowering in long days (58, 59). PFT1 acts as a hub to integrate a variety of interdependent environmental stimuli, including light quality and jasmonic acid-dependent defense responses. An additional gene encodes TERMINAL FLOWER 1 (TFL1), which controls inflorescence meristem identity and regulates flowering time in the long-day flowering pathway. As the QTP experiences a long sunshine duration and a short vegetation growing season, flowering time is particularly critical and affects both the life cycles and reproductive success of many alpine plants (60). In the wild, *C. himalaica* initiates flowering relatively early (April). However, this species rarely blossoms in low-altitude areas outside the QTP. These observations suggest that *C. himalaica* has evolved specific reproductive strategies on photoperiodism and flowering-related pathways as an adaptation to the long sunshine duration and short growing season on the QTP.

**S-Locus Structure and Self-Fertilization of *C. himalaica***. Self-incompatibility in Brassicaceae is controlled by the S-locus recognition system, which prevents self-fertilization (61, 62). Whereas self-fertilization often leads to decreased fitness of homozygous offspring, it also ensures reproduction in absence of pollinators or suitable mates, and therefore can be advantageous for plants to expand their distribution edges and occupy new niches (63, 64). The QTP is characterized by harsh conditions including a short growing season and low pollinator activities. Therefore, it is not surprising that *C. himalaica* is a selfing plant (9). The Arabidopsis/Capsella-like S-locus contains the female specificity gene (S-receptor kinase, SRK) and the male specificity gene (S-locus cysteine-rich protein, SCR) (65). Self-compatibility is achieved when the SRK receptor fails to recognize the SCR ligand, usually through a loss of SCR function (66–69). We manually annotated the S-locus flanking genes (AKR3 and U-box) and the SCR gene on scaffold 29 of the *C. himalaica* genome assembly. Alignment of Crucihimalaya SCR coding sequences with homologs from miscellaneous Arabidopsis halleri, *A. lyrata*, *A. thaliana*, *Capsella grandiflora*, *C. rubella*, and *C. hirsuta* S-locus haplotypes showed that the closest haplotype to *C. himalaica* was the one from A. halleri S15 ([SI Appendix, Fig. S5]) (66, 68, 70, 71).

Note that, due to strong balancing selection on the S-locus, highly diverged S-alleles are shared across species and SCR gene phylogeny does not follow species phylogeny ([SI Appendix, Fig. S5]) (72, 73). SCR protein sequence alignment showed that *C. himalaica* appears to have lost two out of eight conserved cysteines (Fig. 4A) that are essential for structural and functional integrity of SCR ligand (66, 74). This alone may explain the transition to self-fertilization in *C. himalaica*. Interestingly, neither SRK from the S15 haplogroup nor any other SRK had a BLAST hit to the *C. himalaica* genome (Fig. 4B). To investigate whether SRK is indeed lost from the genome, we mapped short reads (450-bp insert size library) of *C. himalaica* to the combination of the *C. himalaica* genome assembly and *A. halleri* S15 haplogroup of S-locus. The SRK gene region did not have any reads mapped, and therefore we conclude that SRK is not present in the *C. himalaica* genome and likely has been lost. Taken together, we reason that the transition to self-compatibility of *C. himalaica* was accomplished by both (i) loss of function in the male recognition gene SCR and (ii) loss of the female recognition gene SRK. It is often suggested that self-fertilization rates of alpine plants increase at higher altitudes (75), due to the reduction in pollinator abundance and seed production in alpine plants (76, 77). This suggestion is consistent with the fact that self-fertilizing populations also exist in *Arabis alpina*, a close relative of *C. himalaica* that grows at high altitudes (78, 79). Moreover, previous studies also found that the rate of self-fertilizing hermaphroditic plants is increasing on the QTP and autonomous self-fertilization provides substantial reproductive assurance in the pollinator scarcity condition of the QTP (80, 81). Therefore, the transition to a self-compatibility mating system in *C. himalaica* likely facilitated its occupation of the QTP.

**Conclusion**

Organisms that live on the QTP face a variety of abiotic stresses under the harsh environmental conditions and presumably have been subjected to a series of adaptive evolutionary changes. In the present study, we de novo-sequenced the genome of *C. himalaica*, a species exclusively found in the Himalayan region. Based on phylogenetic reconstruction and estimated divergence times, we propose that the separation of *C. himalaica* was triggered by the uplift of the QTP, mediated by genomic evolution to adapt to the altered environment. The proliferation of LTR retrotransposons may at least partly be responsible for the increased genome size of *C. himalaica*. The significant contraction and pseudogenization of NBS gene families reflect the strong reduction in pathogen incidence on the QTP. Gene families that underwent significant expansion and genes that show signs of positive selection are enriched in DNA repair and protein ubiquitination pathways, which probably reflect to the adaptation of *C. himalaica* to high radiation, low temperature, and pathogen-depauperate environments on the QTP. Occupation of the QTP by *C. himalaica* was likely facilitated by self-compatibility, which, as we have shown here, involved both male and female components of the recognition system. Both similarities (e.g., DNA repair and disease-resistance pathways) and differences (e.g., reproduction and plant hormone-related pathways) in adaptive mechanisms have been identified among plants and animals that grow at high altitudes. Although further experimental verification is needed, our results provide insights into how plants adapt to harsh and extreme environments.

**Materials and Methods**

**Plant Material, Genome Sequencing, and Assembly.** Seedlings of *C. himalaica* were sampled from Batang County (altitude 4,010 m, N 30.313°, E 99.358°) of QTP. Seedlings from the same individual were cultivated at Kunming Institute of Botany. High-quality genomic DNA was extracted using the Qiagen DNeasy Plant Mini Kit. The *C. himalaica* genome assembly was performed using sequence data obtained from a combination of sequencing technologies: Illumina PE reads, Illumina MP reads, and PacBio RS II reads ([SI Appendix, Table S2]). The whole step of library construction and sequencing was performed at Novogene Bioinformatics Technology Co., Ltd. First, six PE libraries were prepared to sequence the *C. himalaica* genome. These included two PE libraries with insert sizes of 250 and 450 bp and four MP libraries with insert sizes of 2, 5, 10 and 15 kb. Whole genomic sequence (44.49 Gb) data were generated solely using Illumina platforms (HiSEQ 2500) and assembled using ALLPATHS-LG (15). Next, the PacBio reads (6.19 Gb) were used to fill gaps using the PBjelly2 tool (16), followed by scaffold assembly with SPSPACE (82) using long-insert-size PE reads. The accuracy and completeness of the assemblies were assessed by aligning the reads from short-insert-size libraries back to the scaffolds using BWA-mem (v. 0.7.17) (83).

**Gene Prediction and Annotation.** Gene prediction was performed using a combination of homology, de novo, and transcriptome-based approaches. Gene models were integrated by EvidenceModeler ([evidencemodeler.github.io](http://evidencemodeler.github.io)). Gene models were further updated by PASA (84) to generate UTRs and
provide information on alternative splicing variants. The predicted genes were analyzed for functional domains and homologs using InterProScan and BLAST against the NCBI nonredundant protein sequence database, TrEMBL, and SwissProt with an E-value cutoff 1E-15 and Blast2GO with default parameters. Completeness of the genome was also assessed by performing core gene annotation using the BUSCO (17) methods. Orthologous genes among the genomes were clustered into orthologous groups using OrthoMCL (inflation parameter, 1.5) with default parameters. The alignments of each family were concatenated to a super alignment matrix, which was then used for phylogenetic tree reconstruction. We identified orthologous genes among the C. himalaica genome and other related sequences using BLASTP (E value < 1E-7). MCscanX (94) was used to identify syntic blocks within the genome. For each gene pair in a syntenic block, the 4DTv distance was calculated; values less than 20 were retained.

**Whole-Genome Alignment and Duplication Analysis.** We aligned the C. himalaica genome to those of A. thaliana, A. lyrata, and C. rubella using LASTZ (93) with the following parameter values: M = 254 K = 4500 L = 3000 Y = 15000 -seed = match 12 -step = 20 -identity = 85. To avoid the interference caused by repetitive sequences for sequence alignment, RepeatMasker and RepBase library were used to mask repetitive sequences of the above four genomes. The raw alignments were combined into larger blocks using the ChainNet algorithm. We identified orthologous genes among the C. himalaica genome and other related sequences using BLASTP (E value < 1E-7). MCscax (94) was used to identify syntic blocks within the genome. For each gene pair in a syntenic block, the 4DTv distance was calculated; values of all gene pairs were plotted to identify putative whole-genome duplication events and divergence in two species.

**Phylogenetic Tree Construction and Estimation of Species Divergence Times.** We selected genomes of A. thaliana and eight other species (A. thaliana, A. lyrata, C. rubella, Brachypodium distachyon, Zea mays, rice, sorghum, Arabidopsis lyrata, A. halleri, A. lyrata, and C. rubella) for phylogenetic tree construction. We used the Maximum Likelihood (GTR + I + G) model in RAxML software to produce a phylogenetic tree. We used the GTR + I + G model in RAxML software to produce a phylogenetic tree. The tree was further improved by the PROTCATJTT model in RAxML software. We used the GTR + I + G model in RAxML software to produce a phylogenetic tree. The tree was further improved by the PROTCATJTT model in RAxML software. The divergence time between nine species was estimated using MCMCtree in PAML (35) with the options “independent rates” and “GTR” model. A Markov chain Monte Carlo analysis was run for 10,000 generations, using a
burn-in of 1,000 iterations. Three calibration points were applied based on a
substitution rates (σw value). We also deleted all gaps (clean data = 1) from
the alignments to lower the effect of ambiguous bases on the inference of
positive selection. A foreground branch was specified as the clade of C.
himalaica. A likelihood ratio test was conducted to determine whether posi-
tive selection is operating in the foreground branch. In this study, PSGs were
inferred if the P value was less than 0.01. The functional annotation of PSGs in
C. himalaica was also conducted using the same approach and the same
P-value cutoff with the family gene expansion and contraction analysis.

S-Locus Structure and Self-Fertilization of C. himalaica. Known S-haplogroups
from Arabidopsis and Capsella were used as query to search the genome
assembly of C. himalaica (70, 104) using BLASTP. We manually annotated
both flanking genes of the S-locus on scaffold 29: the ARK3 gene and U-box
gene. However, no homologs were detected in our search for the SRK gene.
Using the standard BWA-mem (v. 0.7.17) (83) and Samtools (v. 1.6) (105)
pipeline, we mapped short reads (450-bp insert size library) of C. himalaica
to the combined reference of C. himalaica genome and A. hallieri S15 hap-
logroup of S-locus (70) and then manually explored the coverage of the SRK
region of the S15 haplogroup using igv (v. 2.4.13) (106). The coding sequence
of the predicted SCR gene (ID = AT4G221051.1_At-D2 on scaffold29:1034290-
1034749), together with SCR sequences from A. thaliana, C. rubella, C. grandi-
flora, C. himalaica, C. rubella, was used to design primers and amplifi-
able SCR genes kindly provided by Vincent Castric (Unite Eco-Paléo, CNRS/
Université de Lille 1, Villeneuve d’Ascq, France) from A. hallieri and A. lyrata using
MUSCLE (v. 3.8.31) (97), and a maximum likelihood tree with 500 bootstrap
replicates was constructed using MEGA7 (107) and visualized using Jalview (108).
A comparative structure plot of C. himalaica and A. hallieri S15 S-loci (Fig. 48)
was constructed using the R library genoPlotR (109).

ACKNOWLEDGMENTS. We dearly cherish the memory of our respected
mentor and friend Prof. Yang Zhong at Fudan University for his full support
to this project. We thank Takanori Tsuchimizu and Vincent Castric for their
feedback on the S-locus part. This work was supported by National Natural
Science Foundation of China 31701904, 61872182, 61703157, 61603183, and
31702115; National Key R & D Program of China Grant 2017YF0505200 (to H.S.),
The Strategic Priority Research Program of Chinese Academy of Sciences Grant XDA 20050203 (to H.S.),
National Key Technology Research and Development Program of China Grant 2014AAA020528 (to T.Z.),
the Chinese Academy of Sciences “Light of West China” Program (J.H.), and European Union Seventh Framework Pro-
gramme Grant FP7/2007-2013 under European Research Council Advanced
Grant Agreement 322739 – DOUBLEUP (to Y.V.d.d.S.). P.Y.N. is a postdoctoral
fellow of the Research Foundation-Flanders (122961N).

Assessing assembly and annotation completeness with single-copy orthologs.
Bioinformatics 31:3210–3212.
5. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the
diversity and lineage-specific innovation in vertebrates. Chromosome Res 23:
505–531.
tributors to angiosperm evolution and diversity. Genome Biol Evol 5:1886–1901.
tion of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in
versification and rapid speciation: The case for the Antarctic temperate plant
assembly of the apple genome and gene BIG affect a range of traits in nature. Nature 464:1342-1346.


