The genomes of two Eutrema species provide insight into plant adaptation to high altitudes

Xinyi Guo¹,†, Quanjun Hu¹,†, Guoqian Hao¹,², Xiaojuan Wang¹, Dan Zhang¹, Tao Ma¹, and Jianquan Liu¹,*

¹Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, PR China, and ²Management Committee for Emei Mountain Scenic Area, Biodiversity Institute of Emei Mountain, Leshan 614200, Sichuan, PR China

*To whom correspondence should be addressed. Tel. +028 8541 2053. Fax. +028 8541 2053. Email: liuqj@nwipb.cas.cn

†Equal contributions to this work.

Edited by Dr Satoshi Tabata

Received 16 October 2017; Editorial decision 15 January 2018; Accepted 24 January 2018

Abstract

Eutrema is a genus in the Brassicaceae, which includes species of scientific and economic importance. Many Eutrema species are montane and/or alpine species that arose very recently, making them ideal candidates for comparative studies to understand both ecological speciation and high-altitude adaptation in plants. Here we provide de novo whole-genome assemblies for a pair of recently diverged perennials with contrasting altitude preferences, the high-altitude E. heterophyllum from the eastern Qinghai-Tibet Plateau and its lowland congener E. yunnanense. The two assembled genomes are 350 Mb and 412 Mb, respectively, with 29,606 and 28,881 predicted genes. Comparative analysis of the two species revealed contrasting demographic trajectories and evolution of gene families. Gene family expansions shared between E. heterophyllum and other alpine species were identified, including the disease resistance R genes (NBS-LRRs or NLRs). Genes that are duplicated specifically in the high-altitude E. heterophyllum are involved mainly in reproduction, DNA damage repair and cold tolerance. The two Eutrema genomes reported here constitute important genetic resources for diverse studies, including the evolution of the genus Eutrema, of the Brassicaceae as a whole and of alpine plants across the world.

Key words: Eutrema, high-altitude adaptation, de novo assembly, comparative genomics

1. Introduction

Adaptation to high altitude has contributed to the high level of plant and animal species biodiversity found on numerous mountains across the world.¹ It is well known that animals (including humans) living in highlands have evolved unique genetic adaptations in order to maintain oxygen homeostasis.² In plants, UV-B tolerance and cold tolerance appear to be critical for survival at high altitude.³ However, the genetic mechanisms by which immobile plants adapt to high-altitude conditions remain largely unknown due to a lack of appropriate genome resources.

Species of the Brassicaceae family have a wide range of global distributions and adaptations, as well as different life histories and mating systems, and they therefore comprise an ideal system for many types of study.⁴,⁵ Genome resources are critically important for all these studies.⁶,⁷ Recent advances in high-throughput sequencing technology have facilitated the ability of the scientific community to
sequence the genomes of an increasing number of Brassicaceae species for multiple purposes. 

A previous genome study of an alpine polyploid Brassicaceae species suggested that expansion of numerous gene families was correlated with increased tolerance of cold and of UV-B stress. However, such expansions may have arisen purely from genome duplications rather than because of high-altitude adaptations.

Eutrema is a genus in the Brassicaceae that contains more than 30 species, most of which are distributed in eastern Asia. In addition to the model plant E. salsugineum (Pall.) Al-Shehbaz & Warwick, which is used for studies of abiotic stress, and the commercial crop wasabi, E. japonicum (Miq.) Koidz, the genus contains numerous montane and/or alpine species. Most of these species occur in the Qinghai-Tibet Plateau (QTP) and have originated recently in response to geologic and climatic changes since the Pliocene. In this study, we report the de novo genome sequence assembly and comparative genomic analysis of an alpine plant, E. heterophyllum, from the eastern QTP, and its lowland congener, E. yunnanense, both of which originated from a very recent common ancestor and are diploid perennials with chromosome numbers of 2n = 14 and similar selfing breeding systems, therefore providing an ideal basis for genetic comparisons without the complicating effects of either genome duplication or phyllogenetic evolution. In addition, these two genome assemblies provide important genetic resources for future comparative studies of this genus and family and of high-altitude adaptation in alpine plants.

2. Materials and methods

2.1. Plant materials, genomic DNA and total RNA extraction

We selected and sequenced a pair of Eutrema species that met three criteria: (1) diploid (2n = 14), (2) having diverged recently and (3) growing at contrasting altitudes in their natural distributions (Supplementary Fig. S1). One individual of E. heterophyllum was collected in June 2015 at the Zhuodala pass (~4,700 m above sea level; asl) in Ganzi County, Sichuan Province, China. Fresh and healthy leaves, flowers and root were harvested and immediately frozen in liquid nitrogen, followed by storage at −80 °C in the laboratory prior to DNA/RNA extraction. For comparison, fresh samples of leaf, leafstalk and root tissues were obtained from one E. yunnanense plant collected in April 2014, from Xinning County (~600 m asl), Hunan Province, China. We used the CTAB method to extract total RNA with RNAiso Plus reagent (Takara, Japan) for each tissue collected from both species (i.e. leaves, flowers and roots for E. heterophyllum, and leaves, leafstalks and roots for E. yunnanense). The quality of the extracted RNA was verified by 1% agarose gel electrophoresis. Around 200 μg of genomic DNA was used for library construction. We extracted total RNA with RNAiso Plus reagent (Takara, Japan) for each tissue collected from both species (i.e. leaves, flowers and roots for E. heterophyllum, and leaves, leafstalks and roots for E. yunnanense). The quality of the extracted RNA was verified by 1% agarose gel. DNase treatment was performed before library construction. For each sample, 5 μg of the total high-quality RNA was used for sequencing.

2.2. Genome sequencing and assembly

We constructed Illumina libraries with small (200- to 800-bp) and large (2-, 5-, 10- and 20-kb) inserts and then sequenced them following the Illumina protocols (Illumina, San Diego, CA) using the Illumina HiSeq X Ten platform at Novogene (Tianjin, China). Short reads were first subjected to quality filtering using Trimmomatic v0.33, error correction using BFC (version r181) and mate-pair data deduplication using FastUniq v1.1. We then used SOAPe v2.0116 to estimate the genome size as well as the level of genome-wide heterozygosity, which was based on the 17-mer frequency distribution. We adopted the Platanus17 pipeline, which can efficiently assemble genomes with various levels of heterozygosity. Both genomes were initially assembled into scaffolds using Platanus (version 1.2.2)17 with the first round of gap closing. An additional gap closing procedure was performed with GapCloser (version 1.12). Finally we evaluated genome assemblies for gene coverage using the CEGMA18 and BUSCO (version 2.0)19 pipeline.

2.3. Repeat identification

To structurally annotate repeat sequences in the two Eutrema genomes, we used the TEdenovo pipeline from the REPET package (v2.5) with default parameters. The pipeline performed a self-assembly for the input genome using the BLASTER software (version 2.2.5) to identify and classify repeated elements. In order to attain more comprehensive prediction of TEs, we employed the pipeline to carry out structural detection analysis using LTRharvest (version 1.5.8), RECON v1.0.8, PILER v1.0.24 and GROUPER v2.2.7 to then employ aligned to cluster the matches detected. By performing a multiple sequence alignment using the Map algorithm, a consensus sequence for each cluster was generated to represent the ancestral TE. These consensus sequences were then classified by looking for characteristic structural features and similarities to known TEs in Repbase (20.05.16) and by scanning against the Pfam library with HHMMER. To annotate all TEs across the genome, we further employed the TAnnot pipeline with the default parameters. This pipeline mines the genome sequence using repeated sequences identified in the previous TEdenovo pipeline to produce classified non-redundant consensus repeat sequences along with short simple repeats, which are exported in GFF3 format. We followed the method of Maumus and Quesneville to calculate the age of repetitive DNAs.

2.4. Gene prediction

A combination of ab initio, homology-based and transcript-based methods was used for gene prediction. We first built a comprehensive transcriptome database with the PASA pipeline. After quality control with Trimmomatic v0.33, Illumina RNA-seq reads were assembled into de novo transcripts using Trinity. Then two sets of genome-guided transcripts were built using (1) the genome-guided mode implemented in Trinity and (2) the HISAT-StringTie pipeline. We performed ab initio prediction with Augustus (v3.2.2) using Arabidopsis thaliana as the species model and our comprehensive transcriptome dataset as the input cDNA hint. We used GlimmerHMM (v3.0.4) with the pre-trained directory for A. thaliana genes to generate another set of ab initio predictions. For homology-based prediction, protein sequences of 26,531 E. salsugineum genes as well as the UniRef90 dataset for all Brassicaceae species were aligned to the genome assemblies using SPALN2. Gene models from the three main sources (i.e. aligned transcripts, ab initio predictions and aligned proteins) were merged to produce consensus models by EVIDencerModeler. Weights of evidences for gene models were manually set as: ab initio predictions, weight (Augustus) = 1 and weight (GlimmerHMM) = 1; protein alignments, weight (E. salsugineum) = 2 and weight (Brassicaceae_UniRef90) = 1; transcripts, weight (PASA) = 10. The EVM consensus predictions were updated by performing an additional round of PASA annotation in order to
add UTR and alternatively spliced isoform annotations to gene models.

2.5. Functional annotation
We used BlastP (with E-value ≤ 1 × 10^-5) to assign functional descriptions by carrying out sequence homology searches against different sources, including protein datasets from SwissProt and the Arabidopsis genome annotation version TAIR10. Motifs and domains within gene models were identified by deploying InterProScan (version 5.20.29) with the options -d -g-terms -iprlookup -pathways -p against multiple publicly available databases (Pfam, PROSITE, PRINTS, SMART, TIGRFAMs, Gene3D, PANTHER, etc.). We used AHRD (https://github.com/groundschoof/AHRD (24 January 2018, date last accessed)) to gather information from the search results and added Gene Ontology (GO) information using the GO annotation database for A. thaliana (https://www.ebi.ac.uk/GOA/arabidopsis_release; (24 January 2018, date last accessed)).

2.6. Comparative genomic analysis
We downloaded the protein-coding genes of 13 Brassicaceae species (A. thaliana, A. lyrata, Brassica rapa, Capsella rubella, Leavenworthia alabamica, Sisymbrium irio, Aethionema arboicum, Arabis alpina, Raphanus raphanistrum, Cardamine bursuta, Schrenkella parvula, Thlaspi arvense, E. salsugineum, Aethionema arabicum; see Supplementary Table S4). We used OrthoMCL to delineate gene families and cluster all genes into paralogous and orthologous groups. Gene family expansion and contraction analysis was performed with CAFE3 software. Functional enrichment analysis was carried out by agrGO v2.0. We used R to perform clustering analysis for the gene families of the greatest size in E. heterophyllum. The transformed z-score profile was obtained from the number of genes per species for each family, and the hierarchically clustered (complete linkage clustering) ones, using Pearson correlation as a distance measure. The 2,235 single-copy gene families were used to reconstruct phylogenies with RAxML. Alignment of protein sequences was performed with PRANK and then were translated into coding sequences with PAL2NAL. The MCMCtree program within the PAML package v4.9a was used to estimate neutral diversity. The split of the two major lineages within the Brassicaceae was constrained to be between 20 and 30 million years ago. We used MCScanX to call intra- and inter-species gene collinearity. Analogous substitutions were used to defined gene pairs within identified collinearity blocks using the codeml program implemented in PAML.

2.7. SNP calling and demographic history reconstruction
We used the pairwise sequentially Markovian coalescent (PSMC) model to estimate the history of effective population sizes (Ne) based on genome-wide diploid sequence data. This method has been previously applied to human and other vertebrates, as well as plants, for inferring demographic histories over long evolutionary periods. The quality-filtered Illumina pair end reads with insert sizes of ~500 bp for each species were mapped to the corresponding genome assemblies using BWA (version 0.7.10-r789). For SNP calling, we used the SAMtools (v1.4) pipeline. The settings for PSMC analysis (-p and -t options) were chosen manually according to suggestions given by the authors (https://github.com/Ihhs/psmc (24 January 2018, date last accessed)). We also performed 100 rounds of bootstrapping with the same parameters. The results were combined and plotted using the plot tool in PSMC. A generation time of 3 years and a mutation rate of 7 × 10^-9 per site per year were applied.

2.8. NLR genes and integrated domains
NLR (nucleotide-binding site with leucine-rich repeat) genes in the E. heterophyllum and E. yunnanense genomes were directly extracted from the InterProScan results using the Pfam domain of the nucleotide-binding-adapter shared by APAF-1, R proteins and CED-4 (NB-ARC, PF00931) as query. We further applied the same strategy to identify NLRs in protein sequences from 13 cruciferous species with genomes available. After excluding typical Pfam domains contained in NLRs (i.e. PF01582 for Toll/interleukin-1 receptor [TIR], PF13855 for leucine-rich repeat [LRR] and PF05659 for Resistance to Powdery Mildew 8 [RPW8]), we obtained a list of additional domains that were likely to be fused with NLR proteins. We did not detect TIR2 (PF13676) in our annotated NLRs; this is a TIR domain newly identified among NLR proteins. However, most previously characterized gene fusions were recaptured in this study, indicating that the approach we used for domain searching was conservative and generally reliable. A maximum likelihood tree of phloem protein 2 domains was constructed with RAxML under the PROTGAMMA model with BLOSUM62 matrix, specifying 500 bootstrap replicates. The gene tree was visualized in FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/ (24 January 2018, date last accessed)).

2.9. Positive selection in SCC3 genes
The ratio of non-synonymous substitutions per non-synonymous site (dN) to synonymous substitutions per synonymous site (dS), or ω, is a commonly used indicator of selective pressure acting on protein-coding genes, with ω > 1 representing positive selection, ω = 1 representing neutral evolution and ω < 1 representing purifying selection. To identify signals of positive selection, we extracted SCC3 sequences from all Brassicaceae genomes. A gene tree of aligned protein sequences was constructed using RAxML under the PROTGAMMA model with BLOSUM62 matrix, specifying 500 bootstrap replicates. We used codeml in the PAML software package (v4.9a) to perform the branch-site test among duplicated SCC3 paralogs for a series of pairs of alternative versus null models, aiming to find out whether or not any proportion of coding sites within each paralog has ω > 1 compared with the background level. Log-likelihood values for each pair of nested models were compared using the likelihood ratio test (LRT) to obtain the P-value of significance.

3. Results
3.1. Genome assembly and annotation
After Illumina sequencing and quality control, we obtained 111.87 and 97.29 Gb of Illumina short reads from seven sequencing libraries, corresponding to an estimated 249× and 230× base-pair coverage, for E. heterophyllum and E. yunnanense, respectively (Supplementary Table S1). The estimated genome size of ~400 Mb for both species is larger than that of their congener E. salsugineum, with a higher level of genome-wide heterozygosity in the low-altitude species E. yunnanense (Supplementary Fig. S2). A de novo assembly pipeline allowed us to generate genome assemblies that captured 350 and 413 Mb in 149,415 and 317,771 contigs for the two Extrema genomes, of which 328 and 386 Mb (81% and 91% of the estimated genome sizes) were represented in scaffolds of
1 kb or greater (Table 1). Repetitive DNA constitutes 67% and 70% of, respectively, the *E. heterophyllum* and the *E. yunnanense* genome. Long terminal repeat (LTR) retrotransposons are the most abundant among these repetitive sequences (Supplementary Table S2), having contributed to a recent wave of TE burst (Kimura divergence ≤ 5, Supplementary Fig. S3). Gene completeness analysis revealed that we had achieved more than 95% completeness of gene coverage for both genomes (Supplementary Table S3). We predicted 29,606 genes for *E. heterophyllum* and 28,881 for *E. yunnanense* (Table 1). Clustering of predicted proteomes for 15 Brassicaceae species (Supplementary Table S4) using four-fold degenerate third-codon transversion (4DTv) sites in 2,193 one-to-one orthologous genes identified across their genomes. The nuclear-genome phylogeny (Fig. 1; Supplementary Fig. S4) is congruent with the plastome phylogeny reported previously.51 We dated the divergence of the two species to 3.5 (1.5–7.2) Mya (Fig. 1; Supplementary Fig. S5). Clustering analysis of gene family size revealed that many duplicated genes were retained in genomes known to be polyploid (Fig. 1). We failed to detect any additional whole genome duplication in the two *Eutrema* species when compared with *A. thaliana* using synonymous substitutions per synonymous site (dS) age distribution analysis (Supplementary Fig. S6). We identified 381,883 and 807,391 heterozygous sites in, respectively, the *E. heterophyllum* and the *E. yunnanense* genome. Consistent with the results of analysing K-mer frequency distribution (Supplementary Fig. S2), *E. heterophyllum* had a lower heterozygous SNP rate (1.75 x 10^-3) than *E. yunnanense* (2.18 x 10^-3). A sharp peak was observed in the distribution of genome-wide SNP density for *E. heterophyllum* (Fig. 2A), suggesting that mutations within this genome are more evenly distributed than those in the *E. yunnanense* genome. PSMC analysis based on the local SNP densities of heterozygous sites revealed distinct demographic histories for the two species, in that the high-altitude *E. heterophyllum* showed a decrease-increase-decrease trend while the population size of the low-altitude *E. yunnanense* decreased continuously in the recent past (Fig. 2B).

Table 1. Genome assembly and annotation statistics for two *Eutrema* species

| Assembly statistics | *E. heterophyllum* | *E. yunnanense* |
|---------------------|-------------------|-----------------|
| Estimated genome size (Mb) | 405 | 423 |
| Number of contigs | 149,415 | 317,771 |
| Contig length (Mb) | 350.47 | 412.62 |
| Longest contig (Mb) | 0.67 | 0.39 |
| Contig L50 (seqs) | 1264 | 3269 |
| Contig N50 (kb) | 74.65 | 32.56 |
| Number of scaffolds (>1 kb) | 7,690 | 14,031 |
| Scaffold length (Mb) | 328.16 | 385.58 |
| Longest scaffold (Mb) | 3.5 | 2.07 |
| Scaffold L50 (seqs) | 189 | 341 |
| Scaffold N50 (kb) | 535.32 | 331.5 |

| Annotation statistics | *E. heterophyllum* | *E. yunnanense* |
|-----------------------|------------------|-----------------|
| Number of gene models | 29,606 | 28,881 |
| With InterPro annotations | 22,593 | 22,343 |
| With GO annotations | 20,879 | 20,254 |
| Mean CDS size (bp) | 1201.3 | 1222.3 |
| Mean exon size (bp) | 291.5 | 301.3 |
| Mean intron size (bp) | 233.0 | 210.6 |

Figure 1. Dated phylogeny and gene family analysis of Brassicaceae species. Pie charts showing gains (red) and losses (blue) of gene families are indicated along branches and nodes. The numbers of gene families, orphans (single-copy gene families) and predicted genes, as well as a heat map depicting the hierarchically clustered z-score profile for gene numbers within each family, is indicated next to each species. Background colours on the phylogenetic tree (top to bottom) are basal clade, lineage II and lineage I Brassicaceae species, respectively. Full species names: *Aethionema arabicum*, *Arabis alpina*, *Eutrema yunnanense*, *Eutrema heterophyllum*, *Eutrema saledgesineum*, *Thlaspi arvense*, *Schrenkiella parvula*, *Styosynium irio*, *Raphanus raphanistrum*, *Brassica rapa*, *Cardamine hirsuta*, *Leavenworthia alabamica*, *Capsella rubella*, *Arabidopsis thaliana*, *Arabidopsis lyrata*. The star and triangle denote lineage-specific whole-genome triplications. Mya, million years ago. See online version for color display.
3.3. Expansion of NLR genes and analysis of domain integration

In comparison with other Brassicaceae species, we found that 141 and 178 gene families had significantly expanded \((P < 0.05)\) in the *E. heterophyllum* and *E. yunnanense* genomes, respectively (Supplementary Tables S5–S8). Among the gene families that had expanded within the high-altitude *E. heterophyllum*, GO terms including ‘response to stress’ and ‘nucleotide binding’ \((\text{adjusted } P < 0.05, \text{Supplementary Fig. S7})\) were overrepresented, while no particular GO term was enriched in the case of the low-altitude *E. yunnanense*. The terms enriched in *E. heterophyllum* could be largely attributed to 10 clusters of disease resistance genes encoding nucleotide-binding site with leucine-rich repeats (NLR) proteins (Supplementary Table S6). We noticed that most of these NLR gene families had also expanded in the genomes of three other Brassicaceae species, *A. alpina*, *A. lyrata* and *C. hirsuta*, of which the former two are well documented as alpine plants (Fig. 3). Two NLR families shared across these species account for \(\frac{2}{10}\) of the total number of NLR genes that differ between *E. heterophyllum* and *E. yunnanense* (Supplementary Table S5). Overall, the gene families expanded in *E. heterophyllum* overlap significantly with those expanded in each of these three species \((P < 0.01, \text{Fisher’s exact test})\) (Fig. 3; Supplementary Fig. S8). These results imply that parallel expansion of gene sets, particularly within the NLR genes, may have repeatedly occurred in alpine plants.

Plant NLR genes can fuse with additional domains (hereafter referred to as NB-IDs) that serve as ‘baits’ for pathogen-derived molecules.\(^\text{36}\) We identified 112 NB-IDs with 82 distinct integrated domains across the Brassicaceae genomes (Supplementary Fig. S9; Supplementary Table S9). Within the fused domains, 22 were present in at least two genomes (Supplementary Table S10). Interestingly, in *E. heterophyllum*, we found that integration of one domain encoding phloem protein 2 (PP2, Pfam ID: PF14299), characteristic of a group of plant lectins involved in responses to various stresses,\(^\text{37}\) was shared by two other species, *A. alpina* and *C. hirsuta* (Fig. 4). It is worth noting that the NLRs fused with this domain were clustered into the same gene family that had undergone expansion in all these genomes (Supplementary Table S5). Obviously, the fusion of PP2 arose independently in these three species (Fig. 4), probably in response to the same class of pathogens.

3.4. Species-specific gene duplications and high-altitude adaptation

We also analysed the functional enrichment of genes that were duplicated specifically in the high- or low-altitude species. In the high-altitude *E. heterophyllum*, we observed 30 enriched GO terms within the biological process ontology, including ‘response to stimulus’, ‘response to stress’, ‘reproductive process’ and ‘cell cycle’. However, only four terms within the biological process ontology, ‘developmental process’, ‘anatomical structure development’, ‘multicellular organism development’ and ‘multicellular organismal process’, were
Figure 4. Phylogeny of the phloem protein 2 (PP2) domain fused with expanded NLR genes. Sequences from all *A. thaliana* proteins containing this domain and the integrated PP2 detected in three Brassicaceae genomes (labelled with numbers within circles) were analyzed. Branches are coloured red if a species has a documented high-altitude distribution, and green if not. The structure of NLR fused with PP2 is shown below the gene tree. See online version for color display.

recovered for the low-altitude *E. yunnanense* (Supplementary Tables S11–S14).

A close examination of the families expanded in *E. heterophyllum* revealed multiple genes that are involved in plant reproduction and DNA damage repair (DDR), probably in response to UV-B radiation and other abiotic stresses imposed by the harsh alpine environment. For example, more copies were found for *RPA32*, *SNI1*, and *other abiotic stresses* imposed by the harsh alpine environment.

DNA damage repair (DDR), probably in response to UV-B radiation in the non-homologous end-joining (NHEJ) process (Fig. 5A, Supplementary Table S11). *ICE2* encodes a transcription factor that confers rapid freezing tolerance in *A. thaliana*. The product of FAB1 is involved in membrane lipid synthesis and photosynthesis collapse in the *Arabidopsis fab1* mutant due to degradation of plastids when plants are exposed to low temperature for longer periods. VIN3, which is induced by prolonged winter cold, is essential for the repression of *FLC* during vernalization.

4. Discussion

In this study, we performed *de novo* genome assemblies for *E. heterophyllum* and *E. yunnanense*, a pair of diploid perennials that have diverged recently but have contrasting altitude preferences. Compared with previously published Brassicaceae genomes that were assembled from Illumina short reads, our genome assemblies were of high quality in terms of contiguity and gene coverage, making them suitable for subsequent comparative genomic analysis and thus representing important genetic resources. Phylogenomic analysis confirmed the close relationship and the diploid nature of these two *Eutrema* species. However, the contrasting demographic histories of the two species may reflect their different responses to the local climatic oscillations that occurred during the Quaternary. The dramatic genomic differences observed here may have partly contributed to the high- or low-altitude adaptations that they exhibit.

The NLRs are among the most variable genes across plant species. Variation in these genes between high- and low-altitude *A. thaliana* populations was discovered recently. We found an increase in the number of gene copies for some NLR groups in *E. heterophyllum* and other alpine species of the family Brassicaceae. In addition, we identified a fused domain encoding PP2 among these expanded NLR genes and confirmed the independent acquisition of this domain in different species. According to a recent comprehensive study of fused domains in plant species, this domain also exists in NLRs of Malvaceae and Poaceae species. Thus members of the PP2 family have fused with NLR genes multiple times in different angiosperm lineages. Further studies are needed to investigate whether and how plant immune receptors are involved in high-altitude adaptation in other alpine species.

In the high-altitude *E. heterophyllum*, we found that many genes related to reproduction and DDR had undergone duplications, probably in response to abiotic stress. It should be noted that most of these reproduction-related genes remain as single copies in most angiosperms. In addition, two genes in this category (SCC3 and SMC3) that have undergone duplication events encode cohesin subunits, and similar copy number variations have also been found in animals, where they may cause distinct functional changes in the cohesin complex. To the best of our knowledge, this is the first report of duplication and possible subfunctionalization of cohesin subunits in plants. We have further shown that duplications of SCC3 in *Eutrema* species occurred early, before their divergence. Although we cannot exclude the possibility that the apparent multiple partial SCC3 sequences identified in *E. yunnanense* may be due to the fragmented nature of the draft genome, the evidence for
pseudogenization in this species is strong. The duplication, followed by subsequent loss, of SCC3 in *E. yunnanense* may be correlated with its initial occurrence (or its ancestral distribution) in the high-altitude QTP and subsequent migration to the lower-altitude region. As the cohesin complex has a profound impact on reproduction as well as on gene regulation, duplication of these subunits may have played an important role in plant adaptation to harsh alpine environments. In addition, duplication of genes related to cold acclimation in the high-altitude *E. heterophyllum* may be correlated with plant adaptation to the low and rapidly changing temperatures experienced in alpine habitats.

In summary, we assembled *de novo* genomes of two *Eutrema* species and inferred genetic changes in the alpine species that may underlie high-altitude adaptation, although further tests in more species using multiple approaches are needed. The genome resources that we have developed for the two *Eutrema* species will be very useful for many studies in this genus, the Brassicaceae as a whole, and alpine plants in general.

**Availability**

The short reads of Illumina DNAseq and RNAseq generated in this study are available from the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA) under BioProject PRJNA419026 (BioSample SAMN08200758 for *E. heterophyllum* and SAMN08200758 for *E. yunnanense*). The genome assemblies of this project have been deposited at DDBJ/ENA/GenBank under the accession PKMM00000000 for *E. heterophyllum* and PKML00000000 for *E. yunnanense*.

---

**Figure 5.** Duplication of genes related to DDR in *E. heterophyllum*. (A) Genes involved in the DDR pathway and cell cycle control that are duplicated in *E. heterophyllum*. Duplicated genes are shown in light brown. HR, homologous recombination; NHEJ, non-homologous end-joining. (B) A schematic picture of the plant cohesin complex indicating duplication of genes encoding two cohesin subunits, SCC3 and SMC3. The numbers of the corresponding genes in *E. yunnanense* and *E. heterophyllum* are shown in square brackets in the format [E. yunnanense, E. heterophyllum]. The expression profiles in FPKM (fragments per kilobase per million reads mapped) of paralogs are plotted as log10 values. See online version for color display.

**Figure 6.** Phylogenetic relationships and positive selection of SCC3 genes in Brassicaceae species. Values above branches indicates the support values obtained from 500 bootstrap replicates. Scale bar represents the number of substitutions per site. The branches tested in branch-sites tests of selection for the EhSCC3 and AISC3 genes are indicated. The inset table shows the results of the branch-sites tests, with a P-value based on 1 degrees of freedom. In L, log-likelihood values; H0, null model; H1, alternative model; \( \chi^2 \), chi-square test likelihood ratio (2\*D\*ln L).
Acknowledgements
We thank Bingbing Liu for his help in DNA extraction.

Funding
This work was supported by grants from the National Key Research and Development program (2017YFC0505203), National Natural Science Foundation of China (31590821), National Key Project for Basic Research (2014CB954100), Youth Science and Technology Innovation Team of Sichuan Province (2014TD003), and the 985 and 211 Projects of Sichuan University.

Conflict of interest
The authors declare that they have no competing interests.

Authors’ contributions
J.L. conceived the study. G.H. and X.G. collected samples. X.G. and Q.H. assembled the genomes and performed genome annotation and evolutionary analyses with the help of X.W, D.Z. and T.M. X.G. and J.L. wrote the manuscript. All authors read and approved the final manuscript.

Supplementary data
Supplementary data are available at DNARES online.

References

1. Hughes, C. E. and Atchison, G. W. 2015, The ubiquity of alpine plant radiations: from the Andes to the Hengduan Mountains, New Phytol., 207, 275–82.
2. Beall, M. C. 2014, Adaptation to high altitude: phenotypes and genotypes, Annu. Rev. Anthropol., 43, 251–72.
3. Zhang, J., Tian, Y., Yan, L., et al. 2016, Genome of plant maca (Lepidium meyenii) illuminates genomic basis for high-altitude adaptation in the central Andes, Mol. Plant., 9, 1066–77.
4. Bailey, C. D., Koch, M. A., Mayer, M., et al. 2006, Toward a global phylogeny of the Brassicaceae, Mol. Biol. Evol., 23, 2142–60.
5. Koch, M. A., Dobeš, C., Kiefer, C., Schmickl, R., Klimeš, L. and lysak, M. A. 2006, Supernetwork identifies multiple events of plastid trn F (GA A) pseudogene evolution in the Brassicaceae, Mol. Biol. Evol., 24, 63–73.
6. Koenig, D. and Weigel, D. 2015, Beyond the thale: comparative genomics and genetics of Arabidopsis relatives, Nat. Rev. Genet., 16, 285.
7. Nikiforov, L. A. and Tsiantis, M. 2017, Using mustard genomes to explore the genetic basis of evolutionary change, Curr. Opin. Plant Biol., 36, 119–28.
8. Gan, X., Hay, A., Kwantes, M., et al. 2016, The Cardamine hirsuta genome offers insight into the evolution of morphological diversity, Nat. Plants, 2, 1617.
9. Jiao, W. B., Accinelli, G. G., Hartwig, B., et al. 2017, Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data, Genome Res., 27, 779–86.
10. Willing, E. M., Rawat, V., Mandaková, T., et al. 2015, Genome expansion of Arabidopsis thaliana linked with retrotransposition and reduced symmetric DNA methylation, Nat. Plants, 1, nplants201423.
11. Al-Shewbeh, I. A. and Warwick, S. I. 2005, A synopsis of Eutrema (Brassicaceae), Harvard Pap. Bot., 10, 129–35.
12. Hao, G. Q., Al-Shewbeh, I. A., Ahanì, H., et al. 2017, An integrative study of evolutionary diversification of Eutrema (Eutremaceae, Brassicaceae), Bot. J. Linn. Soc., 184, 204–23.
13. Bolger, A. M., Lohse, M. and Usadel, B. 2014, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics, 30, 2114–20.
14. Li, H. 2015, BFC: correcting Illumina sequencing errors, Bioinformatics, 31, 2885–7.
15. Xu, H., Luo, X., Qian, J., et al. 2012, FastUniq: a fast de novo duplicates removal tool for paired short reads, PLoS One, 7, e52249.
16. Luo, R., Liu, B., Xie, Y., et al. 2012, SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler, Gigascience, 1, 18.
17. Kajitani, R., Toshimoto, K., Noguchi, H., et al. 2014, Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads, Genome Res., 24, 1384–95.
18. Parra, G., Bradnam, K. and Kořík I. 2007, CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes, Bioinformatics, 23, 1061–7.
19. Simao, 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.
20. Flutre, T., Duprat, E., Feuillet, C. and Quesneville, H. 2011, Considering transposable element diversification in de novo annotation approaches, PLoS One, 6, e16526.
21. Quesneville, H., Bergman, C. M., Andrieu, O., et al. 2005, Combined evidence annotation of transposable elements in genome sequences, PLoS Comput. Biol., 1, 166–75.
22. Ellinghaus, D., Kurtz, S. and Willhoeft, U. 2008, LTRharvest, a efficient and flexible software for de novo detection of LTR retrotransposons, BMC Bioinformatics, 9, 18.
23. Bao, Z. and Eddy, S. R. 2002, Automated de novo identification of repeat sequence families in sequenced genomes, Genome Res., 12, 1269–76.
24. Edgar, R. C. and Myers, E. W. 2003, PILER: identification and classification of genomic repeats, Bioinformatics, 21(suppl. 1), i52–8.
25. Huang, X. 1994, On global sequence alignment, Comput. Appl. Biosci., 10, 227–35.
26. Bao, W., Kojima, K. K. and Kohany, O. 2015, Repbase Update, a database of repetitive elements in eukaryotic genomes, Mol. DNA, 6, 11.
27. Finn, R. D., Coggill, P., Eberhardt, R. Y., et al. 2016, The Pfam protein families database: towards a more sustainable future, Nucleic Acids Res., 44, D279–85.
28. Eddy, S. R. 2008, A probabilistic model of local sequence alignment that simplifies statistical significance estimation, PLoS Comp. Biol., 4, e1000069.
29. Maumus, F. and Quesneville, H. 2014, Ancestral repeats have shaped epi-genome and genome composition for millions of years in Arabidopsis thaliana, Nat. Commun., 5, 4104.
30. 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.
31. Grabherr, M. G., Haas, B. J., Yassour, M., et al. 2011, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol., 29, 644–52.
32. Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T. and Salzberg, S. L. 2015, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, Nat. Biotechnol., 33, 290–5.
33. Stanke, M., Dieckmann, B., Baertels, R. and Haussler, D. 2008, Using native and syntactically mapped cDNA alignments to improve de novo gene finding, Bioinformatics, 24, 637–44.
34. Mayoros, W. H., Pertea, M. and Salzberg, S. L. 2004, TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders, Bioinformatics, 20, 2878–9.
35. Yang, R., Jarvis, D. E., Chen, H., et al. 2013, The reference genome of the halophytic plant Eutrema salugenomen, Front. Plant Sci., 4, 1–14.
36. Szupek, B. E., Wang, Y., Huang, H., McGarvey, P. R., Wu, C. H. and UniProt, C. 2015, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches, Bioinformatics, 31, 926–32.
37. Iwata, H. and Gotoh, O. 2012, Benchmarking spliced alignment programs including Spalin2, an extended version of Spalin that incorporates additional species-specific features, Nucleic Acids Res., 40, e161.
