DR. QUANJUN HU (Orcid ID : 0000-0001-6922-2144)

DR. XINYI GUO (Orcid ID : 0000-0001-5416-7787)

PROF. JIANQUAN LIU (Orcid ID : 0000-0002-4237-7418)

Article type : Original Article

Demographic expansion and genetic load of the halophyte model plant *Eutrema salsugineum*

Xiao-Juan Wang[1†], Quan-Jun Hu[1†], Xin-Yi Guo[1], Kun Wang[1], Da-Fu Ru[1], Dmitry A. German[2], Elizabeth A. Weretilnyk[3], Richard J. Abbott[4], Martin Lascoux[5], Jian-quan Liu[1,6*]

[1]Key Laboratory for Bio-resource and Eco-environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610064, China

[2]Department of Biodiversity and Plant Systematics, Centre for Organismal Studies (COS Heidelberg), Heidelberg University, Heidelberg, Germany & South-Siberian Botanical Garden, Altai State University, Barnaul, Russia

[3]McMaster University, Hamilton, Ontario L8S 4K1, Canada

[4]School of Biology, University of St Andrews, St Andrews, Fife KY16 9TH, UK

[5]Department of Ecology and Genetics, Evolutionary Biology Center and Science for Life Laboratory, Uppsala University, Uppsala, Sweden

[6]State Key Laboratory of Grassland Agro-Ecosystem, College of Life Science, Lanzhou University, Lanzhou 730000, Gansu, China

Key words: *Eutrema salsugineum*, migration, hybrid lineage, long-distance dispersal, expansion, deleterious variant

[†]equal contributions to this work; [*]corresponding author (Liujq@nwipb.cas.cn)

Running Title: Demography and genetic load of Salt cress

Abstract

The halophyte model plant *Eutrema salsugineum* (Brassicaceae) disjunctly occurs in temperate to subarctic Asia and North America. This vast, yet extremely discontinuous distribution constitutes an ideal system to examine long-distance dispersal and the ensuing accumulation of deleterious mutations as expected in expanding populations of selfing plants. In this study, we re-sequenced individuals from 23 populations across the range of *E. salsugineum*. Our population genomic data indicate that *E. salsugineum* migrated 'out of the Altai region' at least three times to colonize northern China, northeast Russia and western China. It then expanded its distribution into North America independently from northeast Russia and northern China, respectively. The species colonized northern China around 33.7 thousand years ago (kya) and underwent a considerable expansion in range size approximately 7-8 kya. The western China lineage is likely a hybrid derivative of the northern China and Altai lineages, originating approximately 25-30 kya. Deleterious alleles accumulated in a stepwise manner from (i) Altai to northern China and North America, and (ii) from Altai to northeast Russia and

North America. In summary, *E. salsugineum* dispersed from Asia to North America and deleterious mutations accumulated in a stepwise manner during the expansion of the species' distribution.

Key words: *Eutrema salsugineum*, migration, hybrid lineage, long-distance dispersal, expansion, deleterious variant

Introduction

As predicted by Baker (1955) selfing plants (selfers) show higher colonizing advantages compared to their close outcrossing relatives (outcrossers) and tend to exhibit larger ranges (Randle et al., 2009; Levin, 2010; Grossenbacher et al., 2015). However, the switch to selfing has long been considered as an evolutionary 'dead-end' (Dobzhansky, 1950; Stebbins, 1957) and the observed large range size may be transient and somewhat misleading of the success of a selfing species for two non-exclusive reasons. First, the climatic niche breadth of selfers may decrease more rapidly through time than that of outcrossers because of lowered genetic heterozygosity and decreased effective population size that will likely reduce adaptive potential (Park et al., 2017). Second, genetic load is predicted to be greater in selfers than outcrossers due to reduced purifying selection and lower recombination rates in selfers (Whitlock, 2010; Peterson & Kay, 2015). Indeed, recent studies lend support to the prediction that range expansion and lower effective population size result in the rapid accumulation of deleterious alleles (Lu et al., 2006; Lohmueller, 2014; Mezmouk & Ross-Ibarra, 2014; Renaut & Rieseberg, 2015). For example, in outcrossing poplar (*Populus trichocarpa*), marginal populations exhibit higher levels of deleterious homozygosity than central ones because of recent colonization and smaller effective population sizes (Zhang et al., 2016). Although the accumulation of deleterious alleles should in principle be easier to detect in selfers than outcrossers, since selfers typically expand and migrate through a series of bottlenecks, this prediction has seldom been tested especially during recent

long-distance dispersal events (though see Cao et al., 2011 for evidence of such accumulation in marginal populations of *Arabidopsis*).

The Salt cress (*Eutrema salsugineum* (=*Thellungiella salsuginea*) (Brassicaceae)) (Koch & German, 2013; Wang et al., 2015; Hao et al., 2017) is an "extremophyte" that forms scattered populations across Asia and North America mainly in saline habitats (Inan et al., 2004; Koch & German, 2013; Hao et al., 2017). It occurs most frequently in northern China where it has been collected from numerous localities (Wang et al., 2015), and can be considered as an ideal species for studying the relationship between population expansion and genetic load as its colonization of new areas is expected to have involved strong bottlenecks due to its selfing habit and its specific and narrow ecological requirements. Salt cress is a halophyte showing high tolerance to salt, cold, drought, and oxidative stress and is used as a model species for abiotic resistance studies (e.g. Inan et al., 2004; Griffith et al., 2007; Lamdan et al., 2012; Eshel et al., 2016). Its seeds are similar to those of *Arabidopsis thaliana* in being very small and probably dispersed by wind (Wang et al., 2015). In contrast to the large range of *E. salsugineum*, its two closest relatives, the outcrossing *E. halophilum* and selfing *E. botschantzevii* (Koch & German, 2013) are restricted to northern and central Kazakhstan and adjacent European and Siberian parts of Russia (German, 2006, 2008). How a species with such specific habitat requirements (Inan et al., 2004; Koch & German, 2013; Hao et al., 2017) as those of Salt cress could spread over a large and scattered range on different continents of the Northern Hemisphere remains unclear. Studies of other plant species with disjunct distributions in Asia and North America indicate that they often spread during the early to middle Pliocene via the Bering land bridge which was present for most of the Tertiary until the Pliocene (e.g. Li, 1952; Wen, 1999; Milne, 2006). However, the expansive and disjunct distribution of *E. salsugineum* across Asia and North America might have been established more recently, and likely through long-distance migration (Wang et al., 2015) because populations on both continents show no clear genetic differentiation, especially those from North America and northeast (NE) Russia. Wang et al.'s phylogeographic study, based on 10 nuclear DNA and 9 chloroplast DNA loci, and a recent study based on a single chloroplast DNA fragment (German & Koch, 2017) indicate that North and central American populations of Salt cress were established by multiple long-distance dispersal events from Asia. However, statistical support for

demographic inferences is limited in these two studies and a better characterization of migration routes and timescales is needed to understand changes in genetic load and adaptation in the species.

Here we use whole-genome resequencing of both nuclear and chloroplast genomes to address two specific questions: (i) How and when did *E. salsugineum* establish its current distribution across different continents of the Northern Hemisphere? (ii) How did genome-wide genetic diversity change and deleterious alleles accumulate in the species during its range expansion?

Materials and methods

Sample collection and sequencing

We sampled 92 individuals from 23 sites across the species range (Fig. S1; Tables S1 and S2). The number of individuals sampled per site varied from 15 (Altai, Russia) to one (all four North American populations, three Chinese populations, and one population from Buryatia in Russia). These populations represent most of sites where *E. salsugineum* has previously been reported to occur apart from one recently identified in Central America (German & Koch, 2017). All leaves collected in the field or harvested from individuals grown from germinated seeds were dried and stored in silica gel. In most sites, we sampled different individuals spaced at least 0.5 m. The number of sampled individuals depended on the population size in the field. We sampled only one to two individuals from small populations containing fewer than 10 individuals. In addition, some populations were sampled by non-botanists (especially in North America) and seeds from different individuals were bulked in these cases. If seeds from one site were a mixture from multiple individuals, rather than from separate individuals, only one cultivated individual was selected for analysis. The latitude,

longitude, and altitude of each population were recorded with a GPS, and used to map locations using ArcMap in ArcGIS9.2. We also sampled two closely related species, *E. halophilum* and *E. botschantzevii*, represented by 3 samples each (Fig. S1; Tables S1 and S2).

For each individual, genomic DNA was extracted from leaves using a standard phenol/chloroform procedure (Green & Sambrook, 2012). The integrity and quality of extracted total DNA were checked by monitoring the A260/A280 ratio and by agarose gel electrophoresis. We then constructed paired-end sequencing libraries with an insert size of 600 bp according to the Illumina manufacturer's instructions for sequencing on the Hiseq 2,500 platform and sequenced them using Illumina Hiseq2500 sequencers to generate 2 × 100 bp paired-end reads. The amount of sequence per sample from most locations was targeted to approximately 13× coverage, while there was about 6× coverage per sample from the Xinjiang and Altai populations (Table S2).

Genome mapping and assembly, SNP calling and phylogenetic analyses

After removing low-quality reads using Sickle (https://github.com/najoshi/sickle), sequence data were deposited in the National Center for Biotechnology Information (China) short-read archive (project SRP135200) or BioProject (PRJNA326190). Low-quality reads included PCR duplicates caused by base-calling and adapter contamination, those with average base quality <20, those with >50% having a base quality score <10, and those with >10% unidentified nucleotides. For nuclear genome mapping, all high-quality paired-end reads were aligned to the *Eutrema salsugineum* nuclear genome sequence v1.0 (Salt cress) (Yang et al., 2013)

(http://phytozome.jgi.doe.gov/pz/portal.html), which contains 26,531 predicted protein-coding genes, using BWA software with "mem" option and default parameters (Li & Durbin, 2009; Li et al., 2009). Alignments were further checked for PCR duplicates using PICARD (http://picard.sourceforge.net/). Genome Analysis Toolkit (GATK) (McKenna et al., 2010) was used for base quality recalibrations to enhance alignments in the vicinity of indel polymorphisms. Realignment was performed in two steps: first, RealignerTargetCreator was used to identify regions where realignment was needed; second, IndelRealigner was employed to realign these regions. For each individual, 93.43% of reads were mapped to about 99.07% of the reference genome (Table S2).

After genome mapping, quality filtering of SNP calling was done for all individuals using a Bayesian approach as implemented in SAMtools v1.1 (Li et al., 2009). We used BWA-MEM (Li & Durbin, 2009) to obtain the nuclear SNP datasets with default parameters. Genotype likelihoods and allele frequencies from reads for each individual at each genomic location were also calculated with SAMtools (Li et al., 2009). The 'mpileup' command was used to identify SNPs with the parameters "-C 50 -S -D -m 3 -F 0.002 -q 30 -Q 20 -guf ". Low-coverage depth SNPs (summing all samples) were then filtered with the Perl script vcfutils.pl in BCFtools v1.1 (Li et al., 2009) with parameters "-d 125 -D 1875" and high-quality SNPs (RMS of mapping quality ≥10, the distance of adjacent SNPs in the vicinity of indel polymorphisms ≥5 bp, Hardy-Weinberg equilibrium (HWE) P <0.005, SNP quality ≥30, 3.0≤ quality by depth (each individual) ≤30) were further filtered by Perl scripts (written by XJ or DF) for subsequent analysis. The functional regions (genic, intronic, and intergenic) in which each SNP occurred were determined using annotation information from *E. salsugineum* (Salt cress) annotation (Yang et al., 2013). Genetic SNPs were annotated to be synonymous,

nonsynonymous, 3' and 5' UTR sites using the SnpEff version 4.0 (Cingolani et al., 2012). To identify closely related individuals, the KING program (Manichaikul et al., 2010) was used to estimate the degree of relatedness between all samples based on pairwise comparisons of individual SNP data.

Clean reads from each sample above were also used for chloroplast genome assembly using Velvet software (Zerbino & Birney, 2008) following Guo et al. (2015). The resulting contigs were linked based on overlapping regions after being aligned to the published Salt cress chloroplast genome (*Eutrema salsugineum*, GenBank: KR072661) (Guo et al., 2015) and finally were merged into a consensus linear chloroplast genome sequence using Geneious version 8.0.5 (Kearse et al., 2012). We also added the *Arabidopsis thaliana* chloroplast genome (GenBank: AP000423) as an outgroup before aligning all chloroplast genome sequences using the MAFFT method (Katoh & Standley, 2013).

RAxML was used to construct phylogenetic trees based on the nuclear genome SNPs and chloroplast genomes of all individuals, respectively. We implemented an acquisition bias correction model in the RAxML version 8.0.24 (Stamatakis, 2014) that is intended for analyses of DNA sequences composed exclusively of SNPs. We used a GTR model of nucleotide substitution with a gamma distribution of rate variation across sites for determining the final best tree. Support for each node was assessed based on 1000 rapid bootstrapping replicates. The final Maximum Likelihood trees were viewed using FigTree (v1.4.0) (http://tree.bio.ed.ac.uk/software/figtree/).

Population structure and localized phylogenetic patterns (cacti)

After converting the SNP calling format to a binary ped format using VCFtools (Danecek et al., 2011) and PLINK v1.07 (Purcell et al., 2007), population genetic structure was inferred using two approaches. First, Principal Component Analysis (PCA) was conducted on biallelic SNPs using EIGENSOFT4.2 software (Patterson et al., 2006), with significance levels of eigenvectors determined using the Tracey-Widom test (Patterson et al., 2006). Second, population structure was inferred with the software program ADMIXTURE version 1.23 (Alexander et al., 2009). Simulations for each value of genetic clusters (*K*) ranging from 2 to 10 were run 10 times with 200 bootstrap resampling iterations to estimate the standard error.

To further examine the potential hybrid origin of admixed lineages, we used the HMM-SOM method implemented in Saguaro (Zamani et al., 2013) to identify local phylogenetic relationships of each main lineage based on genome-wide nuclear SNP data. Saguaro was run with default settings to generate 10 different cacti for all populations. *Eutrema halophilum* and *E. botschantzevii* were used as outgroups. FigTree (http://tree.bio.ed.ac.uk/software/figtree/) was employed to view the phylogenetic tree.

Relationships between geographic lineages and demographic history

The relationships between the different geographic lineages and the demographic history of *E. salsugineum* were further inferred using the flexible and robust simulation-based composite maximum likelihood approach based on the site frequency spectrum (SFS) available in the program *fastsimcoal2* (Excoffier et al., 2013). Based on the results of PCA, population

structure and phylogenetic analyses, all sampled materials grouped into four main lineages: northern China, Buryatia and Colorado lineage (Y1), western China lineage (Y2), Altai (Y3) and Canada-NE Russia lineage (Y4). Only loci for which there were data from ≥60 individuals were considered in this analysis. To estimate missing genotypes and to improve genotype accuracy, we used the program *fastPHASE* (Scheet & Stephens, 2006) to infer the haplotypes of all individuals. Genotypes for which posterior probability of the most-likely genotype was less than 0.7 were removed. The final SFS contained 495,798 SNPs. The joint folded SFS of the four groups was used to test different evolutionary scenario parameters to minimize potential biases when presenting the ancestral allelic states. All parameters were estimated from 50 independent fastsimcoal2 runs, with 100,000 coalescent simulations (-n100000, -N100000), as well as 40 ECM cycles (-L40) of the likelihood maximization algorithm under 12 alternative models of phylogenetic relationships and historical events including with or without migration, expansion, hybridization and admixture fitted to the joint SFS (Fig. S2). We assumed a mutation rate of 7.14×10^{-9} per base pair per generation similar to that in *A. thaliana* (Ossowski et al., 2010) and one year per generation was used to convert estimates of divergence time among different populations. The best-fitting model was supported through the maximum value of the likelihood and Akaike Information Criterion (AIC) (Excoffier et al., 2013). Parametric bootstrap estimates were obtained with 50 data sets simulated with estimates of parameters of the best model.

In addition, we used a new statistical tool SMC++ (Terhorst et al., 2017), which combines both a Poisson random field (PRF) model and coalescent HMM approach, to reconstruct the recent history of changes in population size over time for the Y1 lineage using unphased but filtered genome SNPs as input. The Salt cress generation time (g) was set to one year, and neutral mutation rate per generation (μ) was assumed also to be 7.14×10^{-9} per base pair per generation as before. To test for the extensive population expansion of this

species in northern China, we further computed mismatch distributions of the chloroplast genome haplotypes using Arlequin version 3.0 (Excoffier et al., 2005). Haplotype networks were constructed using the Median-Joining model implemented in the program NETWORK 4.0 (Bandelt et al., 1999).

Genetic diversity and linkage-disequilibrium (LD)

Nucleotide diversity was characterized using two standard estimates of the scaled mutation rate: the average pairwise nucleotide diversity, θ_π (Tajima, 1989) and the proportion of segregating sites θ_w (Watterson, 1975) using a window size of 20 kb. We also estimated Tajima's D (Tajima, 1989) as a measure of departure from the standard neutral model and Wright's fixation index (F_{ST}) (Weir & Cockerham, 1984) as an indicator of population differentiation. The above analyses were conducted using Vcftools (vcftools.sourceforge.net), or PERL scripts. For coding regions, we calculated summary statistics for the total coding region, nonsynonymous (N) sites, synonymous (S) sites, intron sites, 3 prime and 5 prime UTR sites, with site identity based on *E. salsugineum* genome annotation (Yang et al., 2013). We also calculated the ratio $\pi N/\pi S$, a measure of purifying selection (Chen et al., 2017), whose logarithmic value is predicted by the nearly neutral theory to be linearly related to the logarithm of πS, used here as a proxy for effective population size (Welch et al., 2008). A high value of the ratio, suggests that purifying selection is inefficient and deleterious mutations accumulate in a species.

We further used the program polydNdS from the libsequence package (Thornton, 2003) to estimate the ratio of nonsynonymous to synonymous difference per synonymous site (dN/dS) (Paape et al., 2013) between *E. salsugineum* and *E.*

botschantzevii through all mapped genes of this species (Yang et al., 2013). Rate of LD decay was estimated using Haploview (Barrett et al., 2005) by calculating pairwise r^2 across all Salt cress individuals or the identified lineage based on SNPs with minor allele frequencies (MAF) greater than 0.05 along 125 scaffolds, each of which was longer than 10 kb. The program fastPHASE (Scheet & Stephens, 2006) with default parameters was used to phase the genotypes into the associated haplotypes. Average r^2 was calculated for pairwise markers in a 500 kb window and averaged across the nuclear genome.

Identification of genetic load

For each individual, a SIFT score predicts whether nonsynonymous SNPs are deleterious or tolerated (Kumar et al., 2009). To classify sites into tolerated and deleterious, protein conservation with homologous sequences and the severity of the amino acid change were considered. If a site is tolerated one expects it to be variable across species whereas if it is deleterious variation will be limited. We used the program SIFT4G (Vaser et al., 2016) to classify nonsynonymous sites into Tolerated and Deleterious and created a specific database for *Eutrema*. To avoid biases when comparing different populations of *E. salsugineum* in subsequent analyses, we used one individual of the closely related species *E. botschantzevii* when classifying nonsynonymous sites into Tolerated and Deleterious (Zhou et al., 2017). Following Zhou et al. (2017), SIFT scores ≤ 0.05 were considered as deleterious and SIFT scores > 0.05 were considered tolerant. SIFT scores range from 0 to 1. We calculated the number of deleterious alleles per individual as 2× homozygous deleterious mutations and heterozygous deleterious mutations (Vaser et al., 2016).

Results

Whole genome sequencing and genetic relatedness

Genome sequence data for each individual ranged from 1.54 to 4.2 Gb, corresponding to sequencing depths of 6 ~ 16× with average genome coverage of 99.07% (Table S2). Across all individuals surveyed, a total of 1,769,662 high-quality SNPs were identified, most of which (81.9%) were located in intergenic regions and a smaller portion (18.1%) within 20,176 genes (Table 1). Among SNPs found within genes, 179,223 (10.13%) were coding SNPs, including 97,254 (5.5%) amino acid replacement nonsynonymous sites (Table 1). Pairwise comparisons of SNPs across individuals detected no duplicate or twin individuals. In addition, a total of 314 SNPs were identified after alignment by MAFFT using all chloroplast genome sequences.

Phylogenetic analyses and population relationships

We conducted phylogenetic analyses of all sampled individuals based on the nuclear genome SNPs and chloroplast genomes. Both datasets clustered all samples into four distinct lineages, Y1 to Y4 (Fig. 1A and B). Y1 comprised all individuals from northern China (Y1-a), one from the southeastern part of Asiatic Russia (Buryatia, Y1-b) and one from North America (Colorado, Y1-c) respectively. Y1-b was placed at the basal position within the Y1 lineage from phylogenetic analyses of both datasets. Y2 contained sampled individuals from western China (Xinjiang) while Y3 contained those from Altai. Y4 comprised all individuals from NE Russia (Y4-a) and Canada (Y4-b). Interrelationships among the four lineages differed greatly between the phylogenetic trees of the two datasets despite high statistical support in both cases. For example, Y3

is sister to Y4 in the nuclear genome phylogeny tree while Y4 is sister to Y1 in the chloroplast genome phylogeny tree.

Principal Component Analysis of the nuclear genome SNPs (Fig. 1C; Table S3) similarly identified the four distinct genetic lineages, and indicated that Y2 was more closely related to Y3, while Y1 and Y4 each comprised two distinct clusters. A network analysis of chloroplast genome haplotypes further distinguished the four main lineages (Fig. 1D) and showed that whereas the Y1-c haplotype from Colorado is derived from the main Y1-a haplotype (from northern China) by only two steps, 15 mutations separate the main Y1-a haplotype from the Y1-b haplotype found in the sample from the southeastern part of Asiatic Russia (Buryatia, Y1-b). ADMIXTURE indicated that the optimum number of genetic clusters (*K*) in the nuclear genome dataset was 2 (i.e., with the lowest cross validation, CV, error, Alexander et al., 2009) (Fig. S3). When *K*= 2, the first cluster (Y1) included individuals from northern China (Y1-a), Colorado of North America (Y1-c) and Buryatia of Russia (Y1-b) while the other cluster comprised individuals from Altai (Y3) and NE Russia-Canada (Y4). Interestingly, individuals from western China (Y2) contained a mixture of the genomes represented by the two main clusters (Fig. 2A), indicative of a hybrid origin. With *K* = 3, the Y4 lineage became evident and separate from the Altai (Y3) lineage while Y2 individuals remained admixed, containing a mixture of the genomes that distinguish the Y1 and Y3 lineages. To test further the hybrid origin of Y2 and estimate the proportions of ancestry contributed by the two putative parental lineages (Y1 and Y3), we used the HMM-SOM method implemented in Saguaro (Zamani et al., 2013) to calculate the local relationships of the four main lineages. We found that across most genomes, Y3 and Y4 were closely related while Y2 comprised a separate lineage (Fig. S4). However, 62% of

the genome of Y2 was closely related to the genomes of Y3, while 21% was closely related to those of Y1 (Fig. S4).

Test of alternative lineage relationships and demographic history

To test further the relationships between the four main lineages (Y1-Y4), we considered 12 alternative models of historical divergence (Fig. S2) and used the coalescent-based composite likelihood estimation method implemented in *fastsimcoal2* (Excoffier et al., 2013) to estimate recent historical demographic parameters of these lineages. We found that model 4 (Fig. S2) was the best fitting demographic model for the nuclear population genetic data (Table S4; Fig. S5). According to this model, the Y1 and Y3 lineages diverged approximately 33 kya years ago from an ancestral population of effective size 73.9 k, while the Y4 lineage (now found in NE Russia and Canada) diverged from Y3 around 28 kya (Fig. 2B and C; Table S4). Approximately 25 to 30 kya the Y2 (western China) lineage originated as a hybrid derivative of the Y1 and Y3 lineages, and most recently (approximately 7.5 kya) the Y1 lineage underwent a three-fold expansion in population size (Fig. 2B; Table S4). A rapid recent expansion in population size of the Y1 lineage was also confirmed by an analysis of the nuclear genomic data using the statistical tool SMC++ (Fig. 2D), and was further indicated by analyses of the network and mismatch distribution for chloroplast genome haplotypes detected in the Y1 lineage (Fig. 2E; Fig. S6).

Genetic diversity and differentiation, minor allele frequency and linkage disequilibrium

Considering the wide geographical range of *E. salsugineum*, its genome-wide average nucleotide diversity was extremely low with θ_w = 0.0018 and θ_π = 0.0014 bp^{-1} (Table 1), while its θπN/θπS ratio was high (0.42, Table 1 and Table S5). Mutation rates and genetic diversity were higher in intergenic than coding regions, and higher in the UTR 5' than in the UTR 3'. Across the genome the number of synonymous mutations was twice as large as the number of nonsynonymous mutations (Table 1). The Y4 lineage had the highest dN/dS ratio and lowest genetic diversity, likely indicating a severe bottleneck and a strong effect of genetic drift, while Y2 had the lowest dN/dS ratio and highest genetic diversity (Table 2), with a positive Tajima's D and high genetic diversity likely reflecting the lineage's recent admixed origin. The dN/dS ratio was lower and genetic diversity and Tajima's D values were higher in Y3 than in the Y1 and Y4 lineages, and the πN/πS ratio was, in general, higher in Y1 and Y4 (with values around 0.40) than in Y2 and Y3 (where it was around 0.35, Table 2).

Linkage disequilibrium (LD) decreased on average to 50% of its initial value at a pairwise distance of about 18 kb at the species level. It increased to 60–70 kb in the northern China (Y1) and the NE Russia-Canadian (Y4) lineages, which is higher than that estimated in the Altai lineage (Y3) (Table 2; Fig. 2G). Genetic differentiation at the genome level between all four lineages (or regions) was estimated to be high (pairwise F*st* = 0.43 ~ 0.72 (Table S6) and AMOVA revealed that 62.26% of the total genetic variation was assigned between lineages with the remainder distributed among populations within regions (Table S6).

Average number of deleterious variants per individual.

SIFT4G (Vaser et al., 2016) was used to predict deleterious nonsynonymous mutations. A total of 33,653 nonsynonymous sites were predicted to be deleterious with most (>90%) present in the homozygous state (Fig. 3). In order to avoid the influence of the different sequence depths among the four main lineages on the number of deleterious variants, we plotted the average ratio of the number of deleterious to synonymous variants for each individual (Liu et al., 2017) (Fig. 3). The mean ratio per individual was found to be lowest in the Altai samples (Y3) (Fig. 3) and significantly higher in the Colorado (Y1-c) and northern China (Y1-a) samples than in the Buryatia (Y1-b) sample. Overall, values for Y1 samples were much higher than those of Y3 samples (Fig. 3). It was also evident that the mean ratio per individual for Canadian (Y4-b) and NE Russia samples (Y4-a) was higher than those for Y3 samples from the Altai, although the stepwise increase in this case, i.e. from Altai to NE Russia and Canada was smaller than the increases from Altai to Buryatia, northern China and Colorado (Fig. 3). The variation trends between different regions for homozygous mutations were consistent with those for the total number (Fig. 3).

Discussion

We used whole-genome sequence data to trace the long-distance dispersal and range expansion of the selfing species, Salt cress (*Eutrema salsugineum*), from Altai to other parts of Asia and into North America, and to assess the link between population expansion and genetic diversity, with particular emphasis on genetic load. Although we sampled only one or two individuals from some populations, we found that such small

samples had a highly similar genetic composition to nearby populations. Thus, though the low number of individuals sampled in some populations may decrease the estimation of genetic diversity, expansion and deleterious alleles in such populations, it should not affect the corresponding values estimated across the regional or total distribution of populations. Our analysis indicated that the species migrated out of Altai three times and colonized North America at least twice during approximately the last 34 kys. Despite the species extensive geographical distribution in the Northern Hemisphere it contains an extremely low level of genome-wide nucleotide diversity. Nonetheless, we found the proportion of deleterious alleles present in samples, increased along both migration routes from Altai to North America, i.e., via northern China to the USA (Colorado), and via NE Russia to Canada, as would be expected if a series of severe bottlenecks, for example, occurred during the colonization process due to long-distance dispersal of few migrants and small effective sizes of populations.

Out of the Altai region, long distance dispersal and inter-lineage hybridization

Eutrema salsugineum diverged from its two close relatives, *E. halophilum* and *E. botschantzevii* in the Altai and probably adjacent regions during the middle Pleistocene between 240 and 480 kya (Al-Shehbaz et al., 2006; Koch & German, 2013; Wang et al., 2015; Guo et al., 2017; Hao et al., 2017). Since its origin, the species has colonised numerous widely separated saline sites in the Northern Hemisphere (Fig. S7) through wind-mediated seed dispersal (Wang et al., 2015). Phylogenetic trees, PCA and ADMIXTURE analysis identified four distinct lineages (Y1-Y4) exhibiting different geographical distributions (Fig. 1 and 2A ; Fig. S7). The Y1 lineage comprises three sub-lineages occurring in populations located in three disjunct regions: Y1-a (in

northern China), Y1-b (Buryatia in Russia) and Y1-c (Colorado in the USA). The Y2 lineage was recovered from one population in western China, close to the distribution of the Y3 lineage in the Altai (Russia), while the Y4 lineage was recovered from populations located in NE Russia and Canada. Within the Y1 lineage, phylogenetic analyses of genome-wide nuclear SNPs and chloroplast genomes, and network analysis of chloroplast haplotypes indicated that sub-lineage Y1-b (in the Buryatia population) was basal (Fig. 1). This sub-lineage is genetically close to the Y3 lineage present in the Altai population (Fig. 1D) and would seem likely, therefore, to provide additional evidence that there was a Altai origin, with sub-lineage Y1-c in Colorado being recently derived (Fig. 1D and 2C). Based on the genetic relationship within the Y4 lineage between populations from NE Russia and Canada, it is possible that the NE Russian population is ancestral to the Canadian one as it is more closely related to the Altai population (Y3) (Fig. 3C). The low level of genome-wide differentiation between the NE Russian and Canadian populations further suggests that the latter population was derived from the former very recently (Fig. 1A and C; Fig. 2A and B). Thus, our phylogenetic analyses indicate that North America was colonized by *E. salsugineum* from Asia on at least two separate occasions, once from NE Russia into Canada and secondly from N China into the USA (Colorado). Very recently, a population of Salt cress was discovered in Central America, and preliminary molecular analysis indicated that it might represent another independent colonization of America by the species from Asia (German & Koch, 2017). However, we were unable to confirm this in the present study, as the population was unknown to us at the time we conducted our genome-wide phylogenetic analyses.

Population genetic analysis of nuclear SNPs indicated that the Y2 lineage found in the western China population, originated from admixture between the Y1-a and Y3 lineages distributed in northern China and the Altai, respectively (Fig. 2A). This was supported by local phylogenetic analyses of the nuclear SNPs across the whole genome (Fig. S4) and confirmed by *fastsimcoal2* tests of alternative models of origin and demographic history (Fig. 2B). However, the chloroplast genome of the Y2 lineage was phylogenetically isolated from both the Y1-a and Y3 lineages, and consequently we could not determine its maternal origin. It is possible that an individual in an unsampled population or one now extinct represents either the Y1-a or Y3 lineage that acted as maternal donor in the origin of the Y2 lineage. A hybrid origin of a distinct lineage of *E. salsugineum* might be considered unexpected given that the species reproduces mainly by selfing (Koch & German, 2013). However, occasional outcrossing occurs in most highly-selfing species and may be relatively common in some populations, for example in *Arabidopsis thaliana* (Luo & Widmer, 2013; Durvasula et al. 2017) where spatial separation of anthers and stigmas within flowers (herkogamy) is correlated with outcrossing rate. In this regard, we noted that some individuals representing the Y3 lineage exhibit herkogamy with stigmas positioned above anthers (Fig. S8).

Genetic diversity and accumulation of deleterious alleles

The very low genome-wide nucleotide diversity recorded in *E. salsugineum* (θ_π = 0.0014 and θ_w = 0.0018) is approximately one quarter of that estimated for *Arabidopsis thaliana* (Nordborg et al., 2005) and one third of the value for the selfing legume *Medicago truncatula* (Branca et al., 2011), but similar to levels of diversity reported for domesticated rice and soybean, both of which experienced domestication bottlenecks

(Table S5). Its very low genetic diversity probably reflects the species recent origin, its high selfing rate, its restriction to highly saline soils (Fig. S7), and the limited size and isolation of the scattered populations it forms (Kirzhner et al., 1996). Our results indicate that the ancestral Y1 lineage was established only 33 kya, with the Y4 and Y3 lineages diverging approximately 28 kya. The demographic expansion of the species in northern China occurred approximately only 7.5 kya and colonization of North America likely took place even more recently. Throughout this very recent range expansion, founder events and genetic drift in small isolated populations formed after long-distance dispersal would have reduced genome-wide nucleotide diversity in areas newly colonized by *E. salsugineum*.

The recent three fold population expansion of Salt cress characterized by repeated genetic bottlenecks complicates the interpretation of the $\theta_\pi N/\theta_\pi S$ ratio as a measure of the accumulation of deleterious mutations, due to relaxed purifying selection. In the present case the overall $\theta_\pi N/\theta_\pi S$ ratio, was 0.42 (Table 1; Table S5), which is particularly high, even for a selfing species (Chen et al. 2017). A high ratio can indicate an accumulation of deleterious mutations and/or that populations have not yet reached equilibrium after a bottleneck. In a nonequilibrium population, a high $\theta_\pi N/\theta_\pi S$ ratio does not necessarily imply that purifying selection is weak, but simply that $\theta_\pi N$ reaches its equilibrium value much faster than $\theta_\pi S$ after a bottleneck (Gordo & Dionisio 2005; Brandvain & Wright 2016) (Table S5). Interestingly both Y1 and Y4 are characterized by extensive LD and negative Tajima's D, suggesting that both populations might indeed be far from equilibrium. Hence, the higher $\theta_\pi N/\theta_\pi S$ ratios observed in these two populations compared to those within Y2 and Y3 may in part be explained by their demographics. This particularly seems to be the case in Y4 where SIFT analysis only

detected a small increase in deleterious alleles compared with that within Y3. However, a much higher ratio of number of deleterious mutations detected by SIFT within Y1 suggests that in this case departure from population equilibrium may play only a limited part in causing a high $\theta_{\pi}N/\theta_{\pi}S$ ratio.

Marginal populations are expected to carry more deleterious mutations than central ones (Lohmueller et al., 2008, 2014; Fu et al., 2014; Zhang et al., 2016; Marsden et al., 2016), particularly in highly-selfing species, which typically expand their ranges through a series of bottlenecks (Chen et al., 2017). In this study, we tested this prediction by comparing the ratio of number of deleterious to synonymous mutations in source and derived populations of Salt cress. In accordance with expectations, an increase in the ratio was found along the proposed migration route from Altai to Buryatia (Russia) (Y1-b) on to northern China (Y1-a) and into the USA (Colorado), while a smaller increase was detected along the other route from northern China to NE Russia, into Canada. In the latter case, a signature of a strong population bottleneck and enhanced genetic drift based on high dN/dS and negative Tajima's D values was evident in the NE Russian and Canadian populations. In these locations, it is feasible that *E. salsugineum* also experiences stronger selection pressure due to a colder climate and higher salinity levels than in other sites (Fig. S7). Differences in selection pressure can lead to differences between populations in the accumulation of deleterious alleles as evidenced for *Arabidopsis* (Gunther & Schmid, 2010). In combination, occasional extreme bottlenecks and strong selection might act to remove some slightly deleterious alleles from a population and this may have occurred during the spread of Salt cress from Asia into Canada leading to the smaller overall increase of deleterious alleles in the NE Russian and Canadian populations.

Of further interest was our finding that the ratio for total (or homozygous) deleterious variants decreased while that for heterozygous deleterious variants increased greatly in the western China population, representing the hybrid lineage (Y2), relative to the composition of populations comprising the parental lineages (Y1 and Y3). It is feasible that hybridization may aid the removal of strongly deleterious alleles or reduce the effect of deleterious ones through heterozygous combinations (Conte et al., 2017).

It has been shown that in selfing *A. thaliana*, more deleterious mutations are present in marginal populations (Cao et al., 2011), while in rice the frequency of deleterious variants increased markedly during domestication due to the hitchhiking effect (Lu et al., 2016). In *E. salsugineum* we found a stepwise increase of deleterious mutations during the migration and expansion of the species. The results of all these studies therefore suggest that the accumulation of deleterious mutations is common during dispersal and migration of selfers. Whereas in outcrossers deleterious mutations mainly result in inbreeding depression (Zhang et al., 2016) that decreases population fitness (Paige 2010; Mezmouk & Ross-Ibarra, 2014), selfers, due to increased homozygosity, should purge deleterious recessive alleles more effectively and experience less inbreeding effects than outcrossers (Paige, 2010). However, the continuous accumulation of deleterious mutations in selfers resulting from demographic effects will reduce their overall fitness, causing for example, decreased germination rates and survival (Keller & Waller, 2002), and also reduced niche breadth (Park et al., 2017). This would explain, at least in part, why selfers are hypothesized to exhibit only 'transient' success after expanding their ranges into new areas often long distances from source populations (Part et al., 2017). Further experiments would be of

value to determine if the phenotypic fitness of *E. salsugineum* is lower in North American populations relative to Altai ones, as expected from the greater accumulation of deleterious alleles within them.

Conclusion

Our population genomic analyses suggest that the sefing species, *E. salsugineum*, migrated from Altai to north China, NE Russia and western China three times and finally colonized North America through two different routes respectively. Along these long-distance migration routes, we examined the accumulation of deleterious mutations and found a stepwise increase of deleterious mutation as expected. However, the level of increase differed according to the colonization route, suggesting that complex demographic and selection effects act on these alleles in natural populations. Despite the fact that *E. salsugineum* has expanded its geographical range over a very wide area during approximately the last 34 kyrs, its accumulating genetic load of deleterious alleles in newly colonized regions may eventually reduce its fitness and ecological niche width leading to a decline in some if not all areas where it is currently found. Future studies of the phenotypic consequences of the deleterious alleles that have accumulated in the species would be highly desirable.

Acknowledgements This work was supported by National Natural Science Foundation of China (31590821, 91331102, 91731301) and the Youth Science and Technology Innovation Team of Sichuan Province (2014TD003). We are highly grateful for

Professor Dirk K. Hincha from Max-Planck-Institut für Molekulare Pflanzenphysiologie, Am Mühlenberg Potsdam D-14476, Germany to kindly provide us some materials from Altai and other regions of Russia. We thank Drs Qiushi Yu and Dongrui Jia for their help in sample collection in China. We thank Huiying Shang, Yazhen Ma, Qian Wang, Hao Bi, Qi He, Qianlong Liang, Zefu Wang, Mingcheng Wang and others in our lab for helps in experiments and analyses.

Availability of data

All Illumina sequence data would be deposited in the National Center for Biotechnology Information short-read archive (project SRP135200; https://www.ncbi.nlm.nih.gov/bioproject/PRJNA326190). Both SNP data and database for deleterious mutations have been deposited at DRYAD doi:10.5061/dryad.6p6k79p.

References

Al-Shehbaz IA, Beilstein MA, Kellogg EA (2006). Systematics and phylogeny of the Brassicaceae (Cruciferae), an overview. *Plant Systematics and Evolution*, **259**, 89–120.

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

Baker HG (1955) Self-compatibility and establishment after 'long-distance' dispersal. *Evolution*, **9,** 347–349.

Bandelt HJ, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, **16**, 37–48.

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Branca A, Paape TD, Zhou P et al. (2011) Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago Truncatula*. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, E864.

Brandvain Y, Wright SI (2016). The limits of natural selection in a nonequilibrium world. *Trends in Genetics*, **32**, 201–210.

Cao J, Schneeberger K, Ossowski S, Gunther T, et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, **43**, 956–963.

Champigny MJ, Sung WW, Catana V et al. (2013) RNA-Seq effectively monitors gene expression in *Eutrema salsugineum* plants growing in an extreme natural habitat and in controlled growth cabinet conditions. *BMC Genomics*, **14**, 578.

Chen J, Glémin S, Lascoux M (2017). Genetic diversity and the efficacy of purifying selection across plant and animal species. *Molecular Biology and Evolution*, **34**, 1417–1428.

Cingolani P, Platts A, Wang LL et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, **6**, 80–92.

Conte GL, Hodgins KA, Yeaman S et al. (2017) Bioinformatically predicted deleterious mutations reveal complementation in the interior spruce hybrid complex:. *BMC Genomics*, **18**, 970.

Danecek P, Auton A, Abecasis G et al. (2011) The variant call format and vcftools. *Bioinformatics*, **27**, 2156–2158.

Dobzhansky T (1950) Evolution in the tropics. *American Scientist*, **38**, 209–221.

Durvasula A, Fulgione A, Gutaker et al. (2017) African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, **114**, 5213–5218.

Eshel G, Shaked R, Kazachkova Y, et al. (2016). *Anastatica hierochuntica*, an *Arabidopsis* desert relative, is tolerant to multiple abiotic stresses and exhibits species-specific and common stress tolerance strategies with its halophytic relative, *Eutrema* (*Thellungiella*) *salsugineum*. *Frontiers in Plant Science*, **7**, 578.

Evanno G, Regnaut S, Goudet J (2010) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L, Dupanloup I, Huertasánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *Plos Genetics*, **9**, e1003905.

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0), an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics*, **1**, 147–150.

Fan K, Wang H (2004). Antarctic oscillation and the dust weather frequency in north China. *Geophysical Research Letters*, **31**, 399–420.

Fu W, Gittelman RM, Bamshad MJ, Akey JM (2014) Characteristics of neutral and deleterious protein-coding variation among individuals and populations. *American Journal of Human Genetics,* **95**, 421–436.

German DA (2006) Additions to the Cruciferae of the flora of Kazakhstan. *Botanicheskii Zhurnal (Moscow & Leningrad)*, **91**, 1198-1211.

German DA (2008) Genus *Thellungiella* (Cruciferae) in Europe. *Botanicheskii Zhurnal (Moscow & Leningrad)*, **93**, 1273-1280.

German DA, Koch MA (2017) *Eutrema salsugineum* (Cruciferae) new to Mexico: a surprising generic record for the flora of middle America. *Phytokeys*, **76**, 13-21.

Gordo I, Dionisio F (2005). Nonequilibrium model for estimating parameters of deleterious mutations. *Physical Review E*, **71**, 031907.

Green MR, Sambrook J (2012) *Molecular Cloning. A Laboratory Manual*. New York: Cold Spring Harbor Laboratory Press.

Griffith M, Timonin M, Wong AC et al. (2007) *Thellungiella*: an *Arabidopsis*-related model plant adapted to cold temperatures. *Plant Cell & Environment*, **30**, 529–538.

Grossenbacher D, Briscoe RR, Goldberg EE, Brandvain Y (2015). Geographic range size is predicted by plant mating system. *Ecology Letters*, **18,** 706–715.

Gunther T, Schmid KJ (2010) Deleterious amino acid polymorphisms in *Arabidopsis thaliana* and rice. *Theoretical Application Genetics*, **121**, 157–168.

Guo XY, Hao GQ, Hu QJ, Wang XJ, Ru DF, Liu JQ (2017) Phylogenetic study of the complete chloroplast genome of Brassicaceae. *BMC Genomics*, **18**,176.

Guo X, Hao G, Ma T (2015) The complete chloroplast genome of Salt cress. *Mitochondrial DNA (Part A)* **27**, 2862-2863.

Hao GQ, Al-Shehbaz IA, Ahani H, Liang QL, Mao KS, Wang Q, Liu JQ (2017) An integrative study of evolutionary diversification of *Eutrema* (Eutremeae, Brassicaceae). *Botanical Journal of the Linnean Society*, **184**, 204–223.

Hey J (2010). Isolation with migration models for more than two populations. *Molecular Biology and Evolution*, **27**, 905–920.

Inan G, Zhang Q, Li P et al. (2004). Salt cress: a halophyte and cryophyte *Arabidopsis* relative model system and its applicability to molecular genetic analyses of growth and development of extremophiles. *Plant Physiology*, **135**, 1718–1737.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Kearse M, Moir R, Wilson A et al. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Keller LF, Waller DM (2002) Inbreeding effects in wild populations. *Trends in Ecology Evolution*, **17**, 230–241.

Kirzhner VM, Korol AB, Nevo E (1996). Complex dynamics of multilocus systems subjected to cyclical selection. *Proceedings of the National Academy of Sciences of the United States of America,* **93**, 6532–6535.

Koch MA, German DA (2013) Taxonomy and systematics are key to biological information: *Arabidopsis*, *Eutrema* (*Thellungiella*), *Noccaea* and *Schrenkiella* (Brassicaceae) as examples. *Frontiers in Plant Science*, **4**, 267.

Kryvokhyzha D, Salcedo A, Eriksson MC et al. (2017) Parental legacy, demography, and introgression influenced the evolution of the two subgenomes of the tetraploid *Capsella bursa-pastoris* (Brassicaceae). *BioRxiv* 234096. doi: https://doi.org/10.1101/234096.

Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, **4**, 1073-1081.

Lamdan NL, Attia Z, Moran N, Moshelion M (2012) The *Arabidopsis*-related halophyte *Thellungiella halophila*: boron tolerance via boron complexation with metabolites? *Plant Cell & Environment*, **35**, 735–746.

Lee YP, Babakov A, de Boer B, Zuther E, Hincha DK (2012) Comparison of freezing tolerance, compatible solutes and polyamines in geographically diverse collections of *Thellungiella sp.* and *Arabidopsis thaliana* accessions. *BMC Plant Biology*, **12**, 131.

Levin DA (2010) Environment-enhanced self-fertilization: implications for niche shifts in adjacent populations. *Journal of Ecology*, **98**, 1276–1283.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A et al. (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.

Li HL (1952). Floristic relationships between eastern Asia and eastern North America. *Transactions of the American Philosophical Society***, 42**, 371–429.

Lohmueller KE, Indap AR, Schmidt S et al. (2008) Proportionally more deleterious genetic variation in European than in African populations. *Nature*, **451**, 994–997.

Lohmueller KE (2014) The distribution of deleterious genetic variation in human populations. Current *Opinion in Genetics & Development*, **29**, 139–146.

Lu J, Tang T, Tang H, Huang J, Shi S, Wu CI (2006) The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends in Genetics*, **22**,126–131.

Luo YH, Widmer A (2013) Herkogamy and its effects on mating patterns in *Arabidopsis thaliana*. *Plos One*, **8**, e57902.

Manichaikul A, Mychaleckyj JC, Rich SS et al. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**, 2867.

Marsden CD, Ortega-Del VD, O'Brien DP et al. (2016) Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 152-157.

Mckenna A, Hanna M, Banks A et al. (2010) The genome analysis toolkit: a map reduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297-303.

Mezmouk S, Ross-Ibarra J. (2014) The pattern and distribution of deleterious mutations in maize. *G3 (Bethesda)*, **4**, 163–171.

Nordborg M, Hu TT, Ishino Y et al. (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *Plos Biology*, **3**, e196.

Ossowski S, Schneeberger K, Lucaslledó JI et al. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, **327**, 92–94.

Paape T, Bataillon T, Zhou PT et al. (2013). Selection, genome-wide fitness effects and evolution ary rates in the model legume *medicago truncatula*. *Molecular Ecology*, **22**, 3525–3538.

Paige KN. (2010) The functional genomics of inbreeding depression: a new approach to an old problem. *BioScience*, **60**, 267–277.

Park DS, Ellison AM, Davis CC (2017). Selfing species exhibit diminished niche breadth over time. *BioRxiv*, doi: http://dx.doi.org/10.1101/157974.

Patterson N, Price AL, Reich D (2006) Population structure and Eigenanalysis. *PLoS Genetics*, **2**, e190.

Peterson ML, Kay KM (2015) Mating system plasticity promotes persistence and adaptation of colonizing populations of hermaphroditic angiosperms. *American Naturalist*, **185**, 28–43.

Purcell S, Neale B, Todd-Brown K et al. (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

Randle AM, Slyder JB, Kalisz S (2009) Can differences in autonomous selfing ability explain differences in range size among sister-taxa pairs of *Eollinsia* (Plantaginaceae)? An extension of baker's law. *New Phytologist*, **183**, 618–629.

Renaut S, Rieseberg LH (2015) The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops. *Molecular Biology and Evolution*, **32**, 2273–2283.

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, **78**, 629–644.

Slotte T, Huang H, Lascoux M, Ceplitis A (2008). Polyploid speciation did not confer instant reproductive isolation in *Capsella* (Brassicaceae). *Molecular Biology and Evolution*, **25,** 1472–1481.

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

Stebbins GL (1957) Self-fertilization and population variability in the higher plants. *American Naturalist*, **91**, 337–354.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology*, **28**, 289–301.

Terhorst J, Kamm JA, Yun SS (2017) Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, **49**, 303–309.

Thornton K (2003). Libsequence: a c++ class library for evolutionary genetic analysis. *Bioinformatics*, **19**, 2325–2327.

Vaser R, Adusumalli S, Leng SN et al. (2016) Sift missense predictions for genomes. *Nature Protocols*, **11**, 1–9.

Wang XJ, Shi DC, Wang XY et al. (2015). Evolutionary migration of the disjunct salt cress *Eutrema salsugineum* (= *Thellungiella salsuginea*, brassicaceae) between Asia and North America. *Plos One*, **10**, e0124010.

Warwick SI, Al-Shehbaz IA, Sauder CA. (2006) Phylogenetic position of *Arabis arenicola* and generic limits of *Aphragmus* and *Eutrema* (Brassicaceae) based on sequences of nuclear ribosomal DNA. *Botany***, 84,** 269–281.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Welch, J. J., Eyre-Walker, A., & Waxman, D. (2008). Divergence and polymorphism under the nearly neutral theory of molecular evolution. *Journal of Molecular Evolution*, **67**, 418–426.

Wen J (1999). Evolution of eastern Asian and eastern north American disjunct distributions in flowering plants. *Annual Review of Ecology & Systematics*, **30**, 421–455.

Whitlock MC (2010) Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. *Evolution*, **54**, 1855-1861.

Wu HJ, Zhang ZH, Wang JY et al. (2012) Insights into salt tolerance from the genome of *Thellungiella salsuginea*. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 12219–12224.

Yang R, Jarvis DE, Chen H et al. (2013) The reference genome of the halophytic plant *Eutrema salsugineum*. *Frontiers in Plant Science*, **4**, 46.

Zamani N, Russell P, Lantz H et al. (2013). Unsupervised genome-wide recognition of local relationship patterns. *BMC Genomics,* **14**, 347–347.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

Zhang M, Zhou L, Bawa R, Suren H, Holliday JA (2016) Recombination rate variation, hitchhiking, and demographic history shape deleterious load in poplar. *Molecular Biology and Evolution*, **33**, 2899–2910.

Zhou Y, Massonnet M, Sanjak JS, Cantu D, Gaut BS (2017) Evolutionary genomics of grape (*Vitis vinifera* ssp. *vinifera*) domestication. *Proceedings of the National Academy of Sciences*, **114**, 11715–11721.

Figure Legends

Fig. 1. Phylogenetic and principal component analyses. (A) and (B) Maximum likelihood (ML) phylogenetic trees using nuclear and chloroplast genome SNPs, respectively. Numbers at each node represent the posterior probability from the Bayesian analysis and the bootstrap values from ML analysis. Y1 including Y1-a: populations from northern China; Y1-b: population from Buryatia; and Y1-c: population from Colorado; Y2: the population from western China (Xinjiang); Y3: the population from Altai; and Y4 including populations from northeastern Russia (Y4-a) and Canada of North America (Y4-b). (C) Principal Component Analysis (PCA) result of the components 1 and 2. (D) Network of all haplotypes excluding singletons detected in chloroplast genomes. Circle size is proportional to haplotype frequency. Number of mutations are indicated on the branches.

Fig. 2. Population structure analysis and demographic history. (A) Population structure inferred by ADMIXTURE with K = 2–5. (B) The best demographic scenario (scenario 4) determined by *fastsimcoal2*. The ancestral population is in gray. The width shows the relative effective population size. The arrows indicate admixture event. (C) The inferred dispersal and colonization history of *E. salsugineum*. The mean predicted geographic distances (km) between the main controversial groups are marked along these possible migration routes. (D) SMC++ analyses of the nuclear genome variations of the populations from northern China (Y1-a lineage). (E) Network relationship of the six plastome (excluding singletons) haplotypes. (F) Distributions of six plastome haplotypes. (G) LD decay of all individuals and four main lineages (X axis stands for physical distances (bp) whereas y axis for r^2).

Fig. 3. The average ratio of number of the deleterious variants (dSNPs) to synonymous variants (sSNPs) in each individual from different lineages or populations predicted in sift4G (Top, the ratio of the total number of dSNPs to sSNPs in each individual; Middle, the ratio of number of homozygotes dSNPs to sSNPs in each individual; Bottom, the proportion of number of heterozygotes dSNPs to the total dSNPs in each individual).

Demographic expansion and genetic load of the halophyte model plant *Eutrema salsugineum*

Xiao-Juan Wang, Quan-Jun Hu, Xin-Yi Guo, Kun Wang, Da-Fu Ru, Dmitry German, Elizabeth A. Weretilnyk, Richard J. Abbott, Martin Lascoux, Jian-quan Liu

Table 1. Coverage and diversity statistics by nucleotide class across the genome for 92 individuals of Salt cress

	Covered bases (Mbp)	Total bases (%)	Polymorphic sites	Total SNPs (%)	θ_π (bp^{-1})	θ_w (bp^{-1})
Total	203	-	1769662	-	0.0014	0.0018
Intergenic	146.8	72.34	1449542	81.91	0.0018	0.0019
Coding	30.1	14.82	179223	10.13	0.0011	0.0011
Nonsynonymous	16.6	8.18	97254	5.50	0.0010	0.0012
Synonymous	4.5	2.21	76590	4.33	0.0024	0.0024
Intron	22.6	11.13	114389	6.46	0.0010	0.0010
UTR 3'	2.7	1.31	17342	0.98	0.0013	0.0013
UTR 5'	0.8	0.41	9166	0.52	0.0021	0.0022

Coding data include replacement, synonymous and other sites such as start-lost and stop-gained ones annotated using snpEff.

Demographic expansion and genetic load of the halophyte model plant *Eutrema salsugineum*

Xiao-Juan Wang, Quan-Jun Hu, Xin-Yi Guo, Kun Wang, Da-Fu Ru, Dmitry German, Elizabeth A. Weretilnyk, Richard J. Abbott, Martin Lascoux, Jian-quan Liu

Table 2. Population genetic summary statistics of the four Salt cress main lineages

Group	the number of individuals	dN/dS	Tajima's D	θ_{π} (bp^{-1})	$\theta_{\pi Nonsyn}$ (bp^{-1})	$\theta_{\pi Syn}$ (bp^{-1})	$\theta_{\pi Nonsyn}/\theta_{\pi Syn}$	LD decays to half of its maximum value (kb)
Y1	63	0.256259	-1.2249	0.000314	0.000340	0.000852	0.40	60.5
Y2	8	0.224750	1.0430	0.000525	0.000323	0.000892	0.34	32.6
Y3	15	0.233234	1.1246	0.000404	0.000218	0.000609	0.36	25.3
Y4	6	0.294439	-1.9732	0.000246	0.000297	0.000727	0.41	67.5

dN/dS was calculated for number of genes (Y1: 20119, Y2: 21700, Y3: 21639 and Y4: 21641 respectively, because dS = 0 for other genes in each group).