Insights into salt tolerance from the genome of *Thellungiella salsuginea*

Hua-Jun Wu[a,1], Zhonghui Zhang[a,1], Jun-Yi Wang[b,1], Dong-Ha Oh[c,1], Maheshi Dassanayake[c,1], Binghang Liu[b,1], Quanfei Huang[b,1], Hai-Xi Sun[a], Ran Xia[a], Yaorong Wu[a], Yi-Nan Wang[a], Zhao Yang[a], Yang Liu[a], Wanke Zhang[a], Huawei Zhang[a], Jinfang Chu[a], Cunyu Yan[a], Shuang Fang[a], Jinsong Zhang[a], Yiqin Wang[a], Fengxia Zhang[a], Guodong Wang[a], Sang Yeol Lee[d], John M. Cheeseman[c], Bicheng Yang[b], Bo Li[b], Jiumeng Min[b], Linfeng Yang[b], Jun Wang[b,2], Chengcai Chu[a,2], Shou-Yi Chen[a,2], Hans J. Bohnert[c,d,e], Jian-Kang Zhu[f,g,2], Xiu-Jie Wang[a,2], and Qi Xie[a,2]

[a]State Key Laboratory of Plant Genomics, National Center for Plant Gene Research (Beijing), Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China; [b]BGI-Shenzhen, Shenzhen 518083, China; [c]Department of Plant Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [d]Division of Applied Life Sciences, Gyeongsang National University, Jinju 660-701, Korea; [e]College of Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia; [f]Shanghai Center for Plant Stress Biology and Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China; and [g]Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, IN 47907

Contributed by Jian-Kang Zhu, June 15, 2012 (sent for review February 27, 2012)

***Thellungiella salsuginea*, a close relative of *Arabidopsis*, represents an extremophile model for abiotic stress tolerance studies. We present the draft sequence of the *T. salsuginea* genome, assembled based on ~134-fold coverage to seven chromosomes with a coding capacity of at least 28,457 genes. This genome provides resources and evidence about the nature of defense mechanisms constituting the genetic basis underlying plant abiotic stress tolerance. Comparative genomics and experimental analyses identified genes related to cation transport, abscisic acid signaling, and wax production prominent in *T. salsuginea* as possible contributors to its success in stressful environments.**

genome sequence | halophyte | gene duplication | stress response

Abiotic stresses such as salinity, drought, or temperature extremes greatly impair plant growth and development and crop yield. The need to cultivate marginal lands to increase food production in the future will expose crops to adverse conditions and exacerbate agricultural problems. Thus, enormous value will come from a better understanding of the mechanisms through which plant tolerance of abiotic stresses is achieved. Most studies on plant response mechanisms leading to stress tolerance have been conducted with the model plant *Arabidopsis*, which has a relatively low capacity to survive abiotic stresses. However, the *Arabidopsis* model and work on a variety of other species have provided clues about enhanced stress tolerance based on individual genes in a number of pathways. Unfortunately, in nearly all cases, genes with a stress-alleviating quality under controlled conditions have failed to generate stress protection in the field. This lack of success argues for developing models that can provide crucial insights into mechanisms that confer high levels of stress tolerance in species that exhibit natural tolerance (1, 2).

The crucifer *Thellungiella salsuginea* (Pallas), a close relative of *Arabidopsis* originally classified as *Thellungiella halophila*, is a halophyte with exceptionally high resistance to cold, drought, and oxidative stresses as well as salinity (1–6). *T. salsuginea* is exemplary by its short life cycle, self-fertility, and being genetically transformable (3). These characteristics make the species an excellent model for unraveling the factors that constitute abiotic stress tolerance (1–8). Further advantages are its relatively small genome size [approximately twice that of *A. thaliana* (3)] and the availability of ecotypes that show a range of stress responses (8).

High-throughput studies of *T. salsuginea* thus far have been restricted largely to the characterization of its transcriptome (5, 7, 8). In addition, comparisons of transcriptome stress responses in *T. salsuginea* and *Arabidopsis thaliana* highlighted different regulation of well-known pathways as well as unstudied stress-related genes (4, 6). The recent publication of the genome sequence of the congeneric species *Thellungiella parvula* has enabled consideration of the genomic and evolutionary basis of stress adaptation with the improved resolution provided by a comparative approach (9).

Here we present the genome sequence and overall chromosome structure of *T. salsuginea* and use comparative genomics and experimental approaches to identify genes in *T. salsuginea* that contribute to its success in stressful environments.

Results

Sequence and Assembly. We sequenced the genome of *T. salsuginea* (Shandong ecotype) using the paired-end Solexa sequencing method (Illumina GA II system). Based on flow cytometry of isolated nuclei stained with propidium iodide (3), we expected a genome size of ~260 Mb (*SI Appendix*, Table S1). Thus, with a total of 34.8 Gb of high-quality sequences, the genome was covered ~134-fold (*SI Appendix*, Table S2). The final length of the assembled sequences amounted to ~233.7 Mb, covering about 90% of the estimated genome size. The assembly consists of 2,682 scaffolds, the 10 longest of which range from 1.9–6.8 Mb (*SI Appendix*, Table S1) and represent 17% of the assembled genome.

In the absence of genetic and physical markers, we assigned many remaining scaffolds to blocks (chromosome segments) identified by Lysak and coworkers (10, 11) by comparative chromosome painting, which represents the ancestral karyotype in the crucifers. By tracing these blocks, we anchored 515 scaffolds onto seven chromosomes, with a total size of 186 Mb (about 80% of the total assembled genome; Fig. 1).

Repetitive Sequences. The size of the *T. salsuginea* genome is approximately twice that of *A. thaliana*, largely reflecting a proliferation of transposable elements (TEs). A repetitive-sequence

Author contributions: J.W., C.C., S.-Y.C., J.-K.Z., X.-J.W., and Q.X. designed research; Z.Z., J.-Y.W., and B.Y. performed research; H.-J.W., Z.Z., D.-H.O., M.D., B. Liu, Q.H., H.-X.S., R.X., Y.Wu, Y.-N.W., Z.Y., Y.L., W.Z., H.Z., J.C., C.Y., S.F., J.Z., Y.W., F.Z., G.W., B. Li, J.M., and L.Y. analyzed data; and H.-J.W., Z.Z., S.Y.L., J.M.C., H.J.B., X.-J.W., and Q.X. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The sequence for the *Thellungiella salsuginea* genome reported in this paper has been deposited in the Data Bank of Japan (DDBJ)/European Molecular Biology Laboratory (EMBL)/GenBank database, http://www.ncbi.nlm.nih.gov/bioproject/?term=txid72664 (accession no. AHIU00000000; PID 80723).

[1]H.-J.W., Z.Z., J.-Y.W., D.-H.O., M.D., B. Liu, and Q.H. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: wangj@genomics.org.cn, ccchu@genetics.ac.cn, sychen@genetics.ac.cn, jkzhu@purdue.edu, xjwang@genetics.ac.cn, or qxie@genetics.ac.cn.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1209954109/-/DCSupplemental.

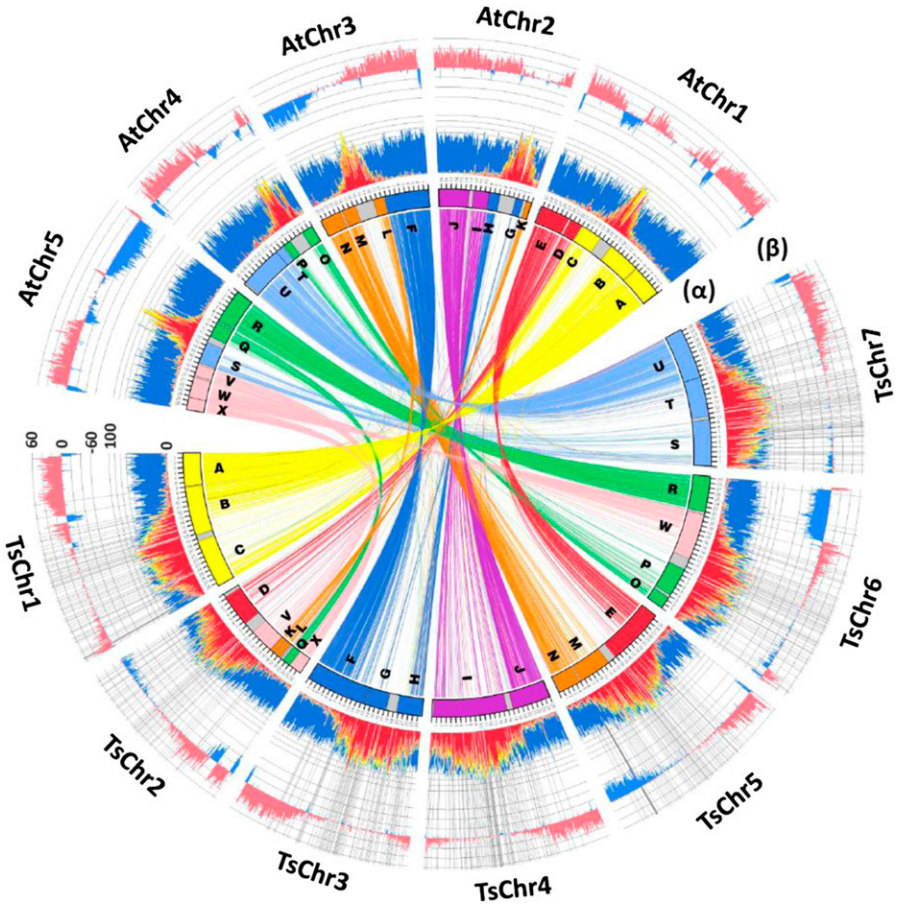


Fig. 1. The genome of *T. salsuginea*. The assembled seven chromosomes of *T. salsuginea* are shown in a comparison with *A. thaliana*. Ancestral karyotype blocks A–X (10) are shown in different colors. Sequences with >70% similarity over the length of 2 kb are connected by links of the same colors as the ancestral karyotype blocks. Histogram α represents the distribution of TEs and predicted genes. Class I retrotransposons, class II DNA transposons, and unclassified repetitive sequences are indicated by red, orange, and yellow colors, respectively, and the predicted genes are shown in blue. The outer histogram β shows the percentage of sequences that can be aligned between the two species with >70% identity. Alignments longer than 500 bp were counted, and their percentages per 100-Kb windows are presented, with the alignments in opposite directions in the two genomes shown in blue and the alignments in the same direction shown in pink. Scales in the *y*-axes of the histograms are in percentage. Radial lines indicate the boundaries of the scaffolds used in the *T. salsuginea* genome assembly.

database search combined with detection of TEs identified 121 Mb of repetitive sequences (*SI Appendix*) representing ~52% of the genome (*SI Appendix*, Tables S1 and S3). This percentage is much higher than the 13.2% and 7.5% TE contents of *A. thaliana* (12) and *T. parvula* (13), respectively. Like most of higher plant genomes, class I TEs (retrotransposons), especially LTR retrotransposons, account for a comparatively high percentage (36%) of the *T. salsuginea* genome. Among these, gypsy and copia are the two most abundant TE families.

Gene Space. A total of 28,457 protein-coding regions were predicted in the sequenced *T. salsuginea* genome using a combination of homologous sequence searches, *ab initio* gene predictions, and transcriptome data comparisons with the genome sequence (*SI Appendix*, Table S1 and Dataset S1). In addition, 447 tRNAs, 11 rRNAs, 432 snRNAs, and 162 microRNAs (including 126 conserved ones) were identified (*SI Appendix*, Tables S1 and S4). The overall ORF length distribution of *T. salsuginea* is comparable to that of *A. thaliana*, with a slightly higher proportion of ORFs shorter than 1,000 bp identified in *T. salsuginea* (*SI Appendix*, Fig. S1). The average exon length of *T. salsuginea* and *A. thaliana* genes is similar (228 and 224 bp, respectively), whereas the average intron length of *T. salsuginea* is ~30% larger than that of *A. thaliana* (200 and 157 bp, respectively) (*SI Appendix*, Table S1) (12).

About 93% of the predicted coding regions showed at least partial similarity with known protein sequences and can be annotated (*SI Appendix*, Table S1). Comparative genomic analysis identified 984 *T. salsuginea* unique gene families and 9,667 families shared by *T. salsuginea*, *A. thaliana*, *Carica papaya*, and *Vitis vinifera* (Fig. 2*A*). Consistent with their close evolutionary relationships, 16,358 gene families were shared by *T. salsuginea* and *A. thaliana*, representing 93.7% and 95.2% of all gene families, respectively (Fig. 2*A*). The protein-coding gene models were compared with *A. thaliana* and *T. parvula* (13), and orthologous gene models were identified. *Thellungiella* species share comparable numbers of orthologs with each other and with *A. thaliana*. Both *Thellungiella* species contain large numbers of "orphan" genes for which no orthologs exist in *A. thaliana* (Fig. 2*B*, indicated by red color). Among all orphan gene models, 54.7%, 62.8%, and 36.5% in *T. salsuginea*, *T. parvula*, and *A. thaliana*, respectively, lacked any Gene Ontology (GO) annotation and hence are annotated as functionally unknown (Dataset S2).

Evolutionary History. Phylogenetic analyses (*SI Appendix*, Fig. S2) indicate a time of divergence between *T. salsuginea* and *A. thaliana* of 7–12 Mya, following the split of the *Arabidopsis* and *Brassica* lineages (14). A similar time has been suggested for the *A. thaliana* and *T. parvula* split (9). Previous studies have suggested that the *A. thaliana* genome shows signatures of the paleohexaploidy whole-genome duplication (WGD) event γ proposed at the base of eudicot divergence and two recent WGD events, β and α, within the crucifer lineage (14). Similarly, two peaks representing the β and α events [~0.28 and ~0.6 fourfold degenerative third-codon transversion (4dTv) distance] were identified in *T. salsuginea* (*SI Appendix*, Fig. S3), suggesting that the divergence of *T. salsuginea* and *A. thaliana* occurred after the two most recent WGD events.

Tandem duplication, segmental duplication, and retrotransposition-directed duplications (*SI Appendix*) were analyzed to weigh their contribution to the variation in gene copy number and to probe for a possible bias in gene functional enrichment in *T. salsuginea* and *A. thaliana*. The total numbers of tandemly

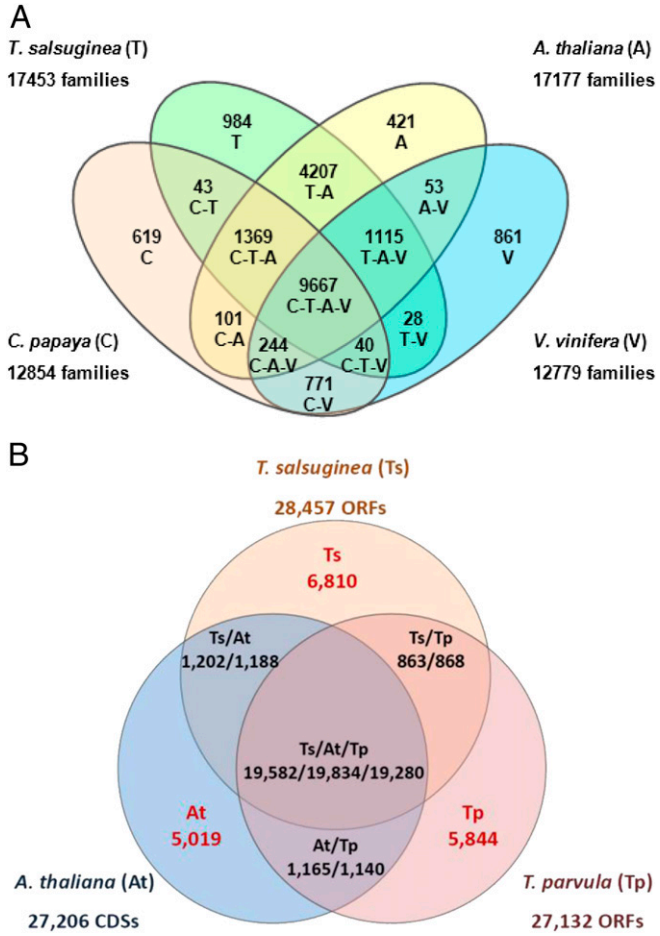


Fig. 2. Comparison of orthologous genes and gene groups. (*A*) Shared orthologous gene clusters among the *T. salsuginea*, *A. thaliana*, *C. papaya*, and *V. vinifera* genomes. Program OrthoMCL was applied to identify orthologous groups among the *T. salsuginea* (T), *A. thaliana* (A), *C. papaya* (C), and *V. vinifera* (V) genomes. (*B*) Shared orthologous genes among crucifers *T. salsuginea* (Ts), *T. parvula* (Tp), and *A. thaliana* (At). Orthologs were identified using OrthoMCL. Genes from different species were considered as orthologs if the shared homology in their deduced amino acid sequences (BlastP, e < 0.00001) was more than 50% of the size of the genes being compared. Numbers of orphan genes lacking an ortholog in the other two species are shown in red. Lists of genes and their GO annotations in each category are given in Dataset S2.

duplicated genes and segmentally duplicated genes are similar in the two species. In contrast, genes carrying an LTR transposon and retrogenes, although they only accounted for a small proportion of genes in total, were significantly more expanded in *T. salsuginea* (Fig. 3 *A* and *B*), probably reflecting the high abundance of TEs, especially LTR retrotransposons. About 30% of the tandemly duplicated genes and 60% of segmentally duplicated genes were shared between *T. salsuginea* and *A. thaliana*. Very few duplication events seemingly mediated by retrogene action, especially those based on LTRs, were conserved, indicating that those events occurred after the divergence of the two species. GO analysis revealed that both tandem and segmental duplications tended to affect genes in similar functional categories (*SI Appendix*, Table S5).

Genome Designed for Stress Response Capacity. GO terms were assigned to the *T. salsuginea* predicted ORFs and *A. thaliana* annotated genes using the Blast2GO pipeline (15). The term "response to stimulus" was identified among the GO categories that differed significantly between *T. salsuginea* and *A. thaliana*, with more genes in the *T. salsuginea* genome classified into this category (Fig. 4). This divergence represents the contribution

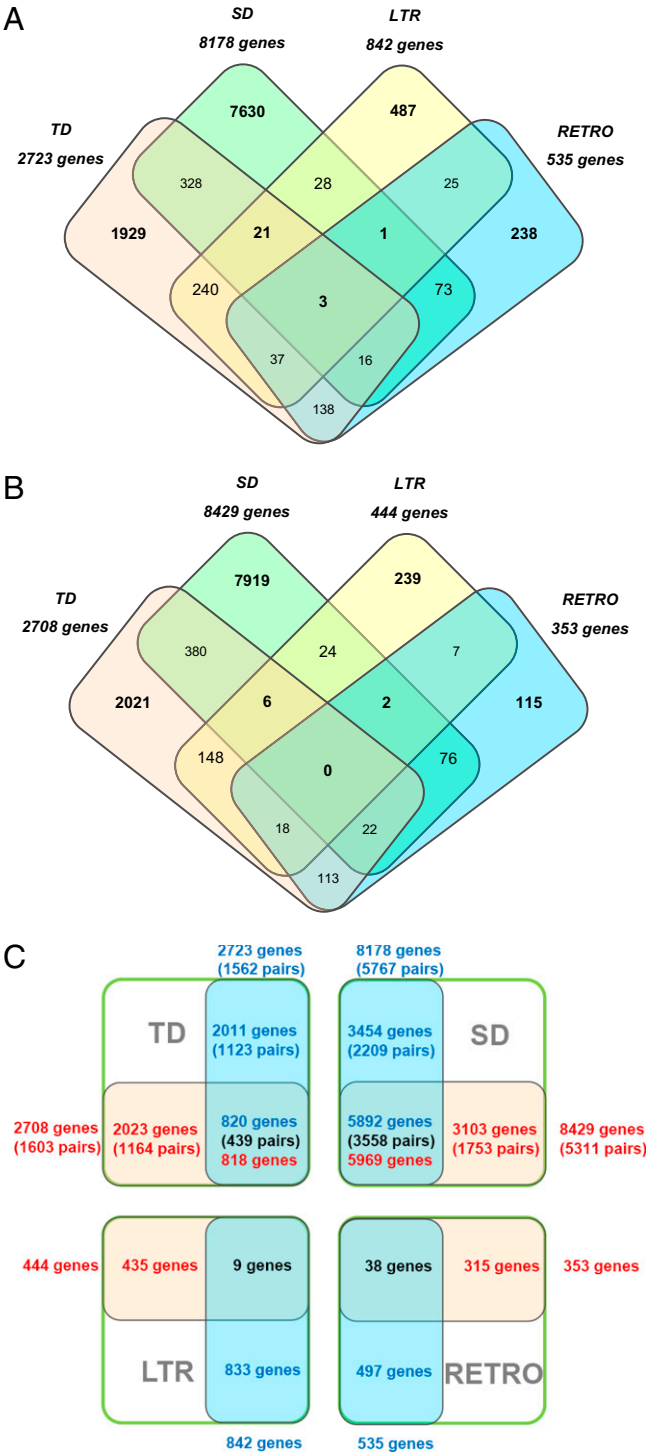


Fig. 3. Comparison of tandem duplication (TD), segmental duplication (SD), and retrotransposition (LTR and RETRO) events in the *T. salsuginea* and *A. thaliana* genomes. (*A* and *B*) Assembled venn plots show shared and specific genes among different types of gene duplications in *T. salsuginea* (*A*) and *A. thaliana* (*B*). (*C*) Comparisons of each type of gene duplication in the two species. Numbers of duplicated genes in *T. salsuginea* are in red; numbers in *A. thaliana* are in blue. LTR, LTR retrotransposon carrying genes; RETRO, retrogenes; SD, segmental duplicated genes; TD, tandem duplicated genes.

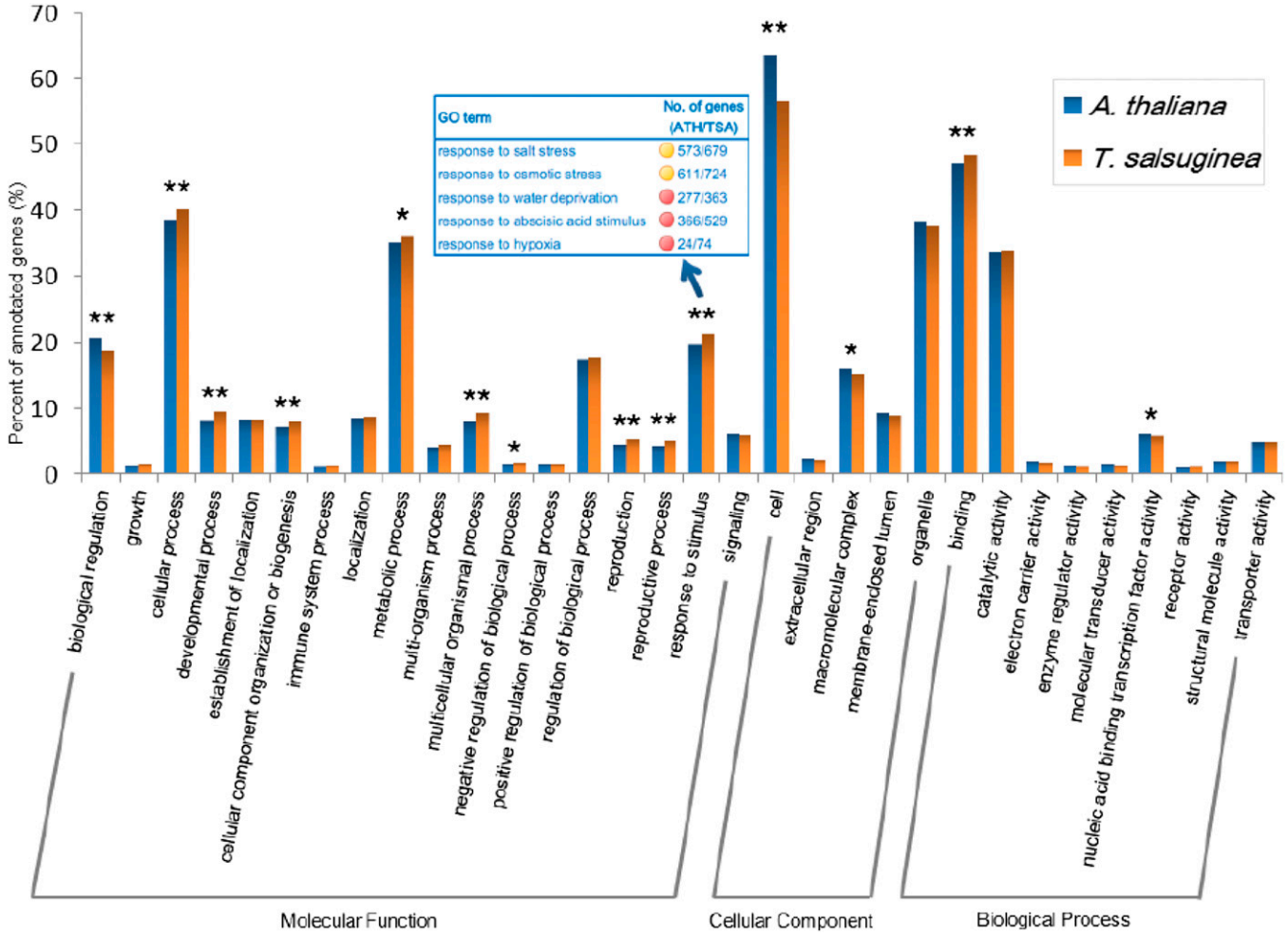


Fig. 4. GO comparison of *T. salsuginea* and *A. thaliana*. Blast2GO results of protein-coding regions from *T. salsuginea* and *A. thaliana* were mapped to categories in the second level of GO terms. Fisher's exact test was used to evaluate the significance of differences in GO category enrichment in the two species. GO terms that contain more than 1% of total genes were included in the graph; those with *P* values below 0.01 and 0.05 are marked by double stars and stars, respectively, on the histogram. Subcategories of the term "response to stimulus" that differ significantly in the two species are shown in the box.

of the combinatory effect of four major duplication events (*SI Appendix*, Table S5). The same trend also was observed in the *T. parvula* genome (13). Detailed analysis revealed that genes related to "response to salt stress," "osmotic stress," "water deprivation," "ABA stimulus," and "hypoxia" were expanded in the "response to stimulus" category in *T. salsuginea* compared with *A. thaliana* (Fig. 4). As a genome signature, this difference may be caused by and could contribute to the high salinity- and drought-tolerant phenotype of *T. salsuginea*.

A total of 21 transcription factor families were found to be expanded in the *T. salsuginea* genome compared with *A. thaliana* (*SI Appendix*, Table S6). These expansions may be associated with the adaptation of *T. salsuginea* to extreme environments, because individual members of some families in *A. thaliana* have been linked previously with stress resistance. For example, the *RAV* gene family, which had been reported to respond to high salt and cold stresses (16, 17), expanded from six members in the *A. thaliana* genome to nine in *T. salsuginea*. Other gene families with known functions in abiotic stress response that expanded in numbers in *T. salsuginea* include the *NF-X1*, *GRAS*, *HSF*, and *Trihelix* families. It has been shown that one *NF-X1* family member, *AtNFXL1*, is required for growth of *Arabidopsis* under salinity stress (18), and *RGL3* in the *GRAS* family can be up-regulated transiently by cold stress (19). *HSFA2*, the most abundant member of the heat-shock response *HSF* family, also is induced by salinity in *Arabidopsis*, and its overexpression enhances salt and osmotic stress tolerance (20). The *GTgamma* subfamily in the *Trihelix* family contains three genes induced by most abiotic stresses in rice (21). Overexpression of two soybean *Trihelix* family genes in *Arabidopsis* greatly enhanced salt, drought, and cold tolerance (22).

Expansion of Genes Related to the Maintenance of Ion Equilibrium. Effective establishment of ionic and osmotic equilibrium is important for plant salinity and drought tolerance. Comparison of gene families involved in ion transport in *T. salsuginea* and *A. thaliana* indicated that gene families providing ionic stress protection, including *HKT, CNGC*, *PPa*, *ACA*, *AVP*, *ATBGL*, *CIPK*, and *CDPK* (23–25), have more members in *T. salsuginea* (*SI Appendix*, Table S7). One group, the *HKT* gene family, encodes Na^+/K^+ transporters that may provide key components affecting or determining salt tolerance in plants (26–29). Recently, two *HKT1* transcripts have been reported in *T. salsuginea* (30); however, the genome annotation revealed a third homolog (*Ts6g08740/TsHKT1;3*). The three *TsHKT1* genes exist in a tandem gene array, similar to the tandem duplication of two *HKT1* genes in *T. parvula* (13); only one copy is present in *A. thaliana* (*SI Appendix*, Table S7). Based on phylogenetic analysis, *Ts6g08650/TsHKT1;1* is clustered with *AtHKT1,* whereas the

other two *TsHKT1* cluster with the two *TpHKT1* genes in another group (*SI Appendix*, Fig. S4*A*). All three *T. salsuginea HKT1* genes were found to be expressed, with the expression of *TsHKT1;2* (*Ts6g08730*) being significantly higher than the expression of the other two genes (*SI Appendix*, Fig. S4*B*).

Stress Tolerance-Supportive Genes and Pathways. Reduction of water loss by epicuticular wax is a strategy used by plants to defend themselves against abiotic stresses (31). Throughout its development, *T. salsuginea* exhibits highly glaucous leaves indicative of complex epicuticular wax organization. We found a tandem duplicated gene in *T. salsuginea* encoding cytochrome P450-dependent midchain hydroxylase MAH1/CYP96A15, which currently is the only known enzyme in the wax-producing–related alkane-forming pathway. This gene is also tandemly duplicated in the *T. parvula* genome (13) but is not duplicated in *A. thaliana* (*SI Appendix*, Fig. S5 and Table S8), perhaps explaining the previous finding that the wax content was much higher in *T. salsuginea* than in *A. thaliana* leaves (32). Genes involved in hormone pathways may serve as another example: The ZEP, AAO, and CYP707A families, all of which are involved in the abscisic acid (ABA) biosynthesis pathway, show an expansion of gene numbers in the *T. salsuginea* genome (*SI Appendix*, Table S9). This expansion may lead to a more complex regulation of ABA biogenesis, contributing to stress tolerance; the induction of gene expression by ABA in *Arabidopsis* is slower than in *T. salsuginea* until much higher stress levels have been reached (4, 6). The rapid ABA response in *T. salsuginea* under high salt conditions may confer a higher salinity-tolerance capacity by slowing down its growth rate. Further experiment evidence is necessary to confirm this hypothesis.

Additional salt stress-related gene families expanded in the *T. salsuginea* genome are summarized in *SI Appendix*, Table S10. Among them, *SAT32* is interesting because it is expanded from one gene in *A. thaliana* and *T. parvula* to six members in the *T. salsuginea* genome (*SI Appendix*, Fig. S6). *AtSAT32* is homologous to human IFN-related developmental regulator (IFRD) and is reported to be involved in salt-stress response (33). It is possible that these expanded family members give *T. salsuginea* more flexibility in response to salinity stress.

Discussion

With the increasing availability of second-generation sequencing, plant genome sequences are appearing in increasing numbers. Because of a desire to understand and improve agronomical important species, crops are an obvious target. A second group includes putative keystone species, i.e., models that might elucidate the evolutionary dimension of the genetic diversity essential to colonization of nearly every climate zone on earth. The third category is plants chosen for their close relationship to existing genomic and genetic models with the goal of expanding potential comparisons relevant at the biochemical and physiological level in particular environments. *T. salsuginea* is such a plant. On the one hand, it is phylogenetically and developmentally similar to the prototypical model, *A. thaliana*. It is a plant with halophytic characters and exceptionally high abiotic stress tolerance, including salinity, cold and freezing temperatures, and the ability to grow in poor soils. During the last decade it also has received considerable attention as a model of physiological and molecular defense against salinity stress (1–8). With the genome sequence presented in this study and with reference to the recently sequenced genome of the congeneric *T. parvula* (13), both juxtaposed with *Arabidopsis*, we have expanded the exploration of gene complement and allele structures that favored extremophile adaptations.

By tracing differences in genome structure and their evolutionary history, we have been able to point to processes that generated two species with extremely divergent adaptations within a time span of 7–12 million years (*SI Appendix*, Fig. S2). Although the gene spaces show extensive colinearity (Fig. 1), and the number of predicted gene models is similar to *A. thaliana* (*SI Appendix*, Fig. S1), selective expansions of seemingly stress-related gene families were observed in the *T. salsuginea* genome (Fig. 4 and *SI Appendix*, Tables S7–S10). Copy number variations of orthologs are largely caused by tandem and segmental duplication events that are unique to each species (Fig. 3*C*), similar to observations in the *T. parvula* genome (13).

However, the *T. salsuginea* genome is characterized by a dramatically higher content of TEs as compared with *A. thaliana* and *T. parvula*, and this greater number of TEs is largely responsible for its enlarged genome size (Fig. 1). Genes contained in LTR and retroelements are more abundant in *T. salsuginea*, with significantly higher numbers of these elements showing tandem duplications than in *A. thaliana* (455 vs. 307 genes). This observation confirms the role of TEs in tandem duplication events (Fig. 3 *A* and *B*). In addition, gene duplications have led to changes in gene dosage. Following sub- and neo-functionalization, functional diversification ensues, and duplicated copies that include favorable characters are retained in the process of natural selection. Duplicates lacking clear advantages for the organism turn into pseudogenes that eventually disappear (34). The stressful environment to which *T. salsuginea* has been exposed seems to have resulted in or to have contributed to the particular population of gene duplications that were retained. This view is supported by the presence of a comparable number but different suite of species-specific duplications in the *A. thaliana* genome (Fig. 3*C*). We also observed a number of translocation events for individual and small groups of genes relative to the *Arabidopsis* and *T. parvula* genomes, although it is not yet possible to assign a particular functional significance to these translocation events. Another outstanding character is the frequency of alterations in the sequences and *cis*-element structures of promoters for orthologous genes in the three species. These alterations can result in a complete rewiring of gene regulation, as exemplified by the expression of the duplicated *HKT1* genes (*SI Appendix*, Fig. S4) as well as for other stress-related genes (9).

Another level of complexity is present in the form of a substantial number of orphan genes that are specific to *T. salsuginea*. These genes do not have an ortholog in *A. thaliana* or in *T. parvula* (Fig. 2*B*) and frequently have no annotation based on sequence similarity. They may represent unique means of adaptation by providing domains with alternative functions, or they may be involved in shuffling of known protein domains. Compared with *Arabidopsis*, *T. salsuginea* is characterized by a dramatically different lifestyle, a unique gene complement, significant differences in the expression of orthologs, and a larger genome size. The *T. salsuginea* genome provides a tool for comparison and contrast to the well-established model, *Arabidopsis*. The resolution provided by the comparison between the two species is exceptionally high. Such resolution is not achievable by comparing plant genomes that are evolutionarily more distant. Multispecies comparative genomics strategies now can focus in detail on gene duplications in stress-related functions, neo-functionalization of duplicated genes, the consequences of translocation events, and orphan gene functions. The divergent regulation of gene expression in development and in communication with stressful environments now can be probed with the support of global transcript profiles. Because fundamental differences in handling stress are emerging (34), it seems that pathways and functions related to stress observed in *Arabidopsis* could be different in evolutionarily stress-tolerant plants. The genome of *T. salsuginea* will be a useful tool in exploring mechanisms of adaptive evolution.

Methods

DNA Library Construction and Sequencing. Short-insert DNA libraries (170 bp, 300 bp, and 800 bp) and long-insert DNA libraries (2 kb, 5 kb, and 10 kb) were built following protocols described previously (35). All libraries were subjected to paired-end sequencing runs, following the manufacturer's user guide (Illumina). A total of 12 DNA libraries were built and sequenced to ensure the randomness of clones. The raw sequence reads with base-calling duplicates, adapter contamination, PCR duplicates, and low-quality sequences were cleaned from the initial sequencing output using custom scripts.

Genome Assembly. We used a hybrid assembly and a hierarchical assembly approach with multiple assembly programs to build gap-free contigs, contigs combined to scaffolds, and scaffolds ordered into pseudochromosomes. At the first level of assembly, we used ABySS (36) and SOAPdenovo (35) followed by minimus2 (37) for meta-assembly of the primary contigs and scaffolds. Contigs were generated with a minimum of 10 overlapping high-quality mate pairs in stringent assembly parameters using 41- to 64-bp-long k-mers in search of high-quality contigs in the primary assemblies. Contigs with lengths less than 100 bp were discarded. In the initial assembly, 50 and 90% of the total length of 174,275,254 bp was covered by contigs larger than 3.23 kb and 149 bp, respectively (contig N50 = 3.23 kb, N90 = 149 bp). Scaffolds were assembled by adding all the paired-end reads to the initial contig assembly, followed by meta-assembly using minumus2. The assembled scaffolds were aligned to both the *T. parvula* (13) and *A. thaliana* (12) genome sequences using Nucmer (38). Scaffolds that could be aligned unambiguously to an ancestral karyotype block (39) in either *T. parvula* or *A. thaliana* were mapped to the karyotype model for the subclade *Eutremae* identified by comparative chromosome painting (10). Directions of mapped scaffolds were visualized further and corrected using the comparative genome visualization tool MAUVE (40), consulting both *T. parvula* and *A. thaliana* genomes. Scaffolds that could not be aligned unambiguously were labeled as unaligned.

More methods and details of data collection are provided in *SI Appendix*.

ACKNOWLEDGMENTS. We thank Beijing Genomics Institute staff members Ying Huang, Na An, Chunfang Peng, Yinqi Bai, Jianwen Li, Qingle Cai, Shiping Liu, Min Xie, Wei Fan, Bo Wang, Sheng Tang, Yuxiang Liu, Juan Wang, Kui Wu, Chuyu Lin, Yalin Huang, Kang Yi, Fei Teng, Fengjie Yu, Haibo Lin, Ruiqiang Li, Zhi Jiang, XiaoJu Qian, Hailong Luo, and Junjie Liu for their sequencing support. This research was supported by Grants 31030047, 30921061, 30825029, and 90917016 from the National Science Foundation of China, Grants 973 2012CB114300 and 2012CB114200 from the National Basic Research Program of China, and by Grant 2009A0714-05 from the State Key Laboratory of Plant Genomics of China. D.-H.O., S.Y.L., and H.J.B. are supported by World Class University Program R32–10148 at Gyeongsang National University, Republic of Korea and the Next-generation BioGreen21 Program SSAC, PJ009030, Rural Development Administration, Republic of Korea.

1. Amtmann A (2009) Learning from evolution: *Thellungiella* generates new knowledge on essential and critical components of abiotic stress tolerance in plants. *Mol Plant* 2:3–12.
2. Bressan RA, et al. (2001) Learning from the *Arabidopsis* experience. The next gene search paradigm. *Plant Physiol* 127:1354–1360.
3. Inan G, 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 Physiol* 135:1718–1737.
4. Taji T, et al. (2004) Comparative genomics in salt tolerance between *Arabidopsis* and *aRabidopsis*-related halophyte salt cress using *Arabidopsis* microarray. *Plant Physiol* 135:1697–1709.
5. Wong CE, et al. (2006) Transcriptional profiling implicates novel interactions between abiotic stress and hormonal responses in *Thellungiella*, a close relative of *Arabidopsis*. *Plant Physiol* 140:1437–1450.
6. Gong Q, Li P, Ma S, Indu Rupassara S, Bohnert HJ (2005) Salinity stress adaptation competence in the extremophile *Thellungiella halophila* in comparison with its relative *Arabidopsis thaliana*. *Plant J* 44:826–839.
7. Taji T, et al. (2008) Large-scale collection and annotation of full-length enriched cDNAs from a model halophyte, *Thellungiella halophila*. *BMC Plant Biol* 8:115.
8. Wong CE, et al. (2005) Expressed sequence tags from the Yukon ecotype of *Thellungiella* reveal that gene expression in response to cold, drought and salinity shows little overlap. *Plant Mol Biol* 58:561–574.
9. Oh DH, et al. (2010) Genome structures and halophyte-specific gene expression of the extremophile *Thellungiella parvula* in comparison with *Thellungiella salsuginea* (*Thellungiella halophila*) and *Arabidopsis*. *Plant Physiol* 154:1040–1052.
10. Mandáková T, Lysak MA (2008) Chromosomal phylogeny and karyotype evolution in x=7 crucifer species (*Brassicaceae*). *Plant Cell* 20:2559–2570.
11. Lysak MA, et al. (2006) Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related *Brassicaceae* species. *Proc Natl Acad Sci USA* 103:5224–5229.
12. Initiative TAG; Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815.
13. Dassanayake M, et al. (2011) The genome of the extremophile crucifer *Thellungiella parvula*. *Nat Genet* 43:913–918.
14. Lysak MA, Koch MA (2011) Phylogeny, Genome, and Karyotype Evolution of Crucifers (*Brassicaceae*). *Genetics and Genomics of the Brassicaceae*, eds Schmidt R, Bancroft I (Springer, New York), pp 1–31.
15. Götz S, et al. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* 36:3420–3435.
16. Fowler SG, Cook D, Thomashow MF (2005) Low temperature induction of *Arabidopsis* CBF1, 2, and 3 is gated by the circadian clock. *Plant Physiol* 137:961–968.
17. Sohn KH, Lee SC, Jung HW, Hong JK, Hwang BK (2006) Expression and functional roles of the pepper pathogen-induced transcription factor RAV1 in bacterial disease resistance, and drought and salt stress tolerance. *Plant Mol Biol* 61:897–915.
18. Lisso J, Altmann T, Müssig C (2006) The AtNFXL1 gene encodes a NF-X1 type zinc finger protein required for growth under salt stress. *FEBS Lett* 580:4851–4856.
19. Achard P, et al. (2008) The cold-inducible CBF1 factor-dependent signaling pathway modulates the accumulation of the growth-repressing DELLA proteins via its effect on gibberellin metabolism. *Plant Cell* 20:2117–2129.
20. Ogawa D, Yamaguchi K, Nishiuchi T (2007) High-level overexpression of the *Arabidopsis* HsfA2 gene confers not only increased themotolerance but also salt/osmotic stress tolerance and enhanced callus growth. *J Exp Bot* 58:3373–3383.
21. Fang Y, Xie K, Hou X, Hu H, Xiong L (2010) Systematic analysis of GT factor family of rice reveals a novel subfamily involved in stress responses. *Mol Genet Genomics* 283: 157–169.
22. Xie ZM, et al. (2009) Soybean Trihelix transcription factors GmGT-2A and GmGT-2B improve plant tolerance to abiotic stresses in transgenic *Arabidopsis*. *PLoS ONE* 4: e6898.
23. Volkov V, Amtmann A (2006) *Thellungiella halophila*, a salt-tolerant relative of *Arabidopsis thaliana*, has specific root ion-channel features supporting K^+/Na^+ homeostasis under salinity stress. *Plant J* 48:342–353.
24. Sun ZB, et al. (2008) Overexpression of a *Thellungiella halophila* CBL9 homolog, ThCBL9, confers salt and osmotic tolerances in transgenic *Arabidopsis thaliana*. *J Plant Biol* 51(1):25–34.
25. Lv S, et al. (2008) Overexpression of an H^+-PPase gene from *Thellungiella halophila* in cotton enhances salt tolerance and improves growth and photosynthetic performance. *Plant Cell Physiol* 49:1150–1164.
26. Vera-Estrella R, Barkla BJ, García-Ramírez L, Pantoja O (2005) Salt stress in *Thellungiella halophila* activates Na^+ transport mechanisms required for salinity tolerance. *Plant Physiol* 139:1507–1517.
27. Rus A, et al. (2006) Natural variants of AtHKT1 enhance Na^+ accumulation in two wild populations of *Arabidopsis*. *PLoS Genet* 2:e210.
28. Ren ZH, et al. (2005) A rice quantitative trait locus for salt tolerance encodes a sodium transporter. *Nat Genet* 37:1141–1146.
29. Byrt CS, et al. (2007) HKT1;5-like cation transporters linked to Na^+ exclusion loci in wheat, Nax2 and Kna1. *Plant Physiol* 143:1918–1928.
30. Ali Z, et al. (2012) TsHKT1;2, a HKT1 homolog from the extremophile *Arabidopsis* relative *Thellungiella salsuginea*, shows $K(^+)$ specificity in the presence of NaCl. *Plant Physiol* 158:1463–1474.
31. Kosma DK, et al. (2009) The impact of water deficiency on leaf cuticle lipids of *Arabidopsis*. *Plant Physiol* 151:1918–1929.
32. Teusink RS, Rahman M, Bressan RA, Jenks MA (2002) Cuticular waxes on *Arabidopsis thaliana* close relatives *Thellungiella halophila* and *Thellungiella parvula*. *Int J Plant Sci* 163(2):309–315.
33. Park MY, et al. (2009) Isolation and functional characterization of the *Arabidopsis* salt-tolerance 32 (AtSAT32) gene associated with salt tolerance and ABA signaling. *Physiol Plant* 135:426–435.
34. Oh DH, Dassanayake M, Bohnert HJ, Cheeseman JM (2012) Life at the extreme: Lessons from the genome. *Genome Biol* 13:241.
35. Li R, et al. (2010) The sequence and *de novo* assembly of the giant panda genome. *Nature* 463:311–317.
36. Simpson JT, et al. (2009) ABySS: A parallel assembler for short read sequence data. *Genome Res* 19:1117–1123.
37. Sommer DD, Delcher AL, Salzberg SL, Pop M (2007) Minimus: A fast, lightweight genome assembler. *BMC Bioinformatics* 8:64.
38. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.
39. Schranz ME, Lysak MA, Mitchell-Olds T (2006) The ABC's of comparative genomics in the Brassicaceae: Building blocks of crucifer genomes. *Trends Plant Sci* 11:535–542.
40. Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 14:1394–1403.

Supporting Information Appendix

Insights into Salt Tolerance from the Genome of *Thellungiella salsuginea*

Hua-Jun Wu[1]*, Zhonghui Zhang[1]*, Jun-Yi Wang[2]*, Dong-Ha Oh[4]*, Maheshi Dassanayake[4]*, Binghang Liu[2]*, Quanfei Huang[2]*, Hai-Xi Sun[1], Ran Xia[1], Yaorong Wu[1], Yinan Wang[1], Zhao Yang[1], Yang Liu[1], Wanke Zhang[1], Huawei Zhang[1], Jinfang Chu[1], Cunyu Yan[1], Shuang Fang[1], Jinsong Zhang[1], Yiqin Wang[1], Fengxia Zhang[1], Guodong Wang[1], Sang Yeol Lee[5], John M Cheeseman[4], Bicheng Yang[2], Bo Li[2], Jiumeng Min[2], Linfeng Yang[2], Jun Wang[2,†], Chengcai Chu[1,†], Shou-Yi Chen[1,†], Hans J Bohnert[4,5], Jian-kang Zhu[3,†], Xiu-Jie Wang[1,†] and Qi Xie[1,†]

[1] State Key Laboratory of Plant Genomics, National Center for Plant Gene Research (Beijing), Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Datun Road, Beijing 100101, China

[2] BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China

[3] Department of Horticulture & Landscape Architecture, Purdue University, West Lafayette, Indiana 47907, USA

[4]Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[5]Division of Applied Life Sciences, Gyeongsang National University, Jinju 660-701, Korea

[6]College of Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia

* These authors contributed equally to this work.

[†]Corresponding authors

To whom correspondence should be addressed: E-mail: qxie@genetics.ac.cn or xjwang@genetics.ac.cn or zhu132@purdue.edu

FAX: 86-10-64889351; 86-10-64860941

Table of Contents

Supporting Materials and Methods

Assembly accuracy

The accuracy of the assembled genome was confirmed using available ESTs and BAC sequences. Nearly 98% of all ESTs showed exact sequence matches with the assembled genome over at least 50% of their entire length. Four BAC sequences from NCBI and two from BGI showed 95% coverage and greater than 99.9% accuracy of low repeat regions.

Repeat annotation

Known TEs were identified using RepeatMasker (version 3.3.0) to search against the Repbase TE library (version 15.11) (1). TEdenovo pipeline included in the REPET (2) package was used for identifying novel repetitive sequences. Default parameters were used except for "minNbSeqPerGroup: 5". The resultant *de novo* output identified consensus TEs, excluding sequences classified as "NoCat", was used as the reference repeats library in a second RepeatMasker run to identify and mask novel repetitive sequences in the *T. salsuginea* genome.

Gene prediction and annotation

Protein coding gene models were identified by FGENESH++ pipeline (Softberry Inc., Mount Kisco, NY) with parameters trained with *A. thaliana* gene models. Genome sequences masked by RepeatMasker using RepBase and the *de novo* reference TE library as described in Repeat annotation section were used as input. To facilitate the gene prediction with transcriptome evidence, a *T. salsuginea* reference transcriptome was assembled from Illumina RNA-seq reads using Abyss and Vmatch (http://www.vmatch.de/). Known *T. salsuginea* ESTs and full-length cDNA sequences from NCBI database were added to the reference transcriptome. *De novo* predicted gene models were corrected based on comparison to all known plant protein sequences from the NCBI NR database. The reference transcriptome was aligned to the genome sequence and used to identify the borders of exons and untranslated regions (UTRs) for gene models with transcriptome evidence. Open reading frame (ORF) sequences less than 150 nucleotides were filtered out. The nucleotide ORF and protein sequences were annotated based on sequence homology to known sequences, using BlastN and

BlastP (E-value ≤ 1e-5) to search against the NCBI nt and nr databases (ftp://ftp.ncbi.nih.gov/blast/db/), respectively. The Blast2GO pipeline was used for Gene Ontology annotation, with the incorporation of InterProScan and KEGG pathway search results (3).

Gene family analysis

We used a best hit strategy for systemic identification of gene copy number variations in gene families in *T. salsuginea*. All *T. salsuginea* genes were subjected to BlastP search (E-value ≤ 1e-5) against all *A. thaliana* genes. The best hit to each *T. salsuginea* gene were picked up and considered as its most close orthologous gene in *A. thaliana*. A gene relationship table was generated based on the best hit strategy and was then used to calculate the gene copy number variations in each collected family. Transcription factor gene families in *A. thaliana* were downloaded from PlantTFDB (4), and stress related gene families in *A. thaliana* were manually collected from published records. Gene family member variations in other species were performed similarly. For comparison of gene models with *A. thaliana* and *T. parvula*, protein-coding gene models in TAIR10 (www.arabidopsis.org) and the version 2.0 annotation of *T. parvula* (www.thellungiella.org) were used. Gene models were clustered using OrthoMCL. Orthologous gene pairs were defined as sharing deduced amino acid sequence homology (BlastP, E-value < 1e-5) over 50% of the total length of the shorter gene being compared.

Identification of segmental and tandem duplications

To identify segmental duplications, we first performed self BlastP (-v 5 -b 5 -e 1e-10) using the deduced protein sequences of the *T. salsuginea* and *A. thaliana* genomes. A Perl script provided by DAGchainer was used to remove the repetitive matches (5). This was done by clustering all groups of matched genes that fall within 50 kb of each other and reporting only the single highest scoring match in each region. Segmental duplicated blocks were then identified using DAGchainer with optimized parameters (-s -I -D 200000 -g 10000 for *A. thaliana*; -s -I -D 500000 -g 25000 for *T. salsuginea* because of the large number of transposon insertions). To identify tandem duplications, we performed self BlastP using protein sequences with the parameters -v 100 -b 100 -e 1e-5. All genes were grouped with the following parameters: identity ≥ 70%; coverage ≥ 30%. Homologous genes within the same group and with fewer than five genes in between were identified

as tandem duplicated gene pairs.

LTR retrotransposon carrying genes and retrogenes

We used a similar method to that described by Jiang *et al.* (6) to perform systemic identification of LTR retrotransposons carrying genes and retrogenes. Full-length LTR retrotransposons were identified by using LTR_FINDER (7) with parameters -S 5 -C, which will contain at least 5 of 11 typical structural or sequence features of LTR retrotransposons. Protein coding genes entirely located within these LTR retrotransposons were considered as LTR (retrotransposon) carrying genes. To find retrogenes, we performed BlastP using the single-exon protein sequences as query, multiple-exon protein sequences as database and used the cutoff of identity ≥ 70%, query coverage ≥ 70 % and E value < 1e-8 to select retrogenes.

Phylogenetic tree construction and species divergent time estimation

The phylogenetic tree of the *T. salsuginea* and the other plant genomes was constructed using the 2226 single-copy orthologuous genes and 4-fold degenerate sites (4dTv) method. The divergence time between *T. salsuginea* and *A. thaliana* was estimated by the MULTIDIVTIME program.

Quantification of TsHKT1 transcripts with real time reverse transcription polymerase chain reaction (RT-PCR)

RNA samples from *A. thaliana* and *T. salsuginea* seedlings with and without salt stress were prepared essentially as described by Oh et al (8). To deduce absolute copy numbers of transcripts per μg total RNA samples, calibration curves were generated by performing real time PCR using 7900 HT Fast Real-Time PCR system (Applied Biosystems, Carlsbad, CA) with serial dilutions of known amount of recombinant plasmid DNA molecules that contain the template sequences (9). The recombinant plasmids were prepared by cloning RT-PCR products amplified by the following primers into the pGemTeasy vector (Promega, Madison, WI):

AtHKT1 223F GAAGTCTTCTCCAACACCCAACTT

AtHKT1 823R TACTTGAGGGATTAGGAGCCAGA

TsHKT1;1 44F TTGCTAAAAATCCTTCCGTCCTCT

TsHKT1;1 770R CCCGAAACGAGAAACAATAAAAAGC

TsHKT1;2 409F AATCATGTCAAGCTTTCTAGTCAG

TsHKT1;2 1152R TCCTTTAATTTCATCTCCGGAATCGTGT

TsHKT1;3 424F GATCATGTCAAGATTTCTAGTCAGA

TsHKT1;3 1181R AAATCCACTTTTCTTTCCCTTCTTTTCATTTC

Real time RT-PCR was performed using primers that are specific to each of the *A. thaliana* and *T. salsuginea HKT1* gene homologs. From the real time RT-PCR results and the calibration curves, the absolute transcript copy numbers were calculated as described by Pfaffl (9). Primer sequences are listed below:

AtHKT1 476F CGGTGGTTCTTAGTTACCATCTT

AtHKT1 594R GAGAGGTGAGATTTCTTTGGAACT

TsHKT1;1 195F GTCTCCTCCATGTCCACCATCG

TsHKT1;1 305R AGAGTGTGAGGAATGAAGTAAAGACCTCG

TsHKT1;2 782F CAAATCGAGAAGAATTGGGTTACATTCT

TsHKT1;2 903R GCAGAATAGAAGAAACTGTATCATCACAAGC

TsHKT1;3 785F CAAAGCGCGACGAATTTGGTTATATTC

TsHKT1;3 928R GCAGAAGAGAAGAAACTGTATCATCACAAAC

References

1. Jurka J, *et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110:462-467.

2. Flutre T, Duprat E, Feuillet C, & Quesneville H (2011) Considering transposable element diversification in de novo annotation approaches. *PLoS One* 6:e16526.

3. Gotz S, *et al.* (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* 36:3420-3435.

4. Zhang H, *et al.* (2011) PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Res* 39:D1114-1117.

5. Haas BJ, Delcher AL, Wortman JR, & Salzberg SL (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* 20:3643-3646.

6. Jiang SY, Christoffels A, Ramamoorthy R, & Ramachandran S (2009) Expansion mechanisms and functional annotations of hypothetical genes in the rice genome. *Plant Physiol* 150:1997-2008.

7. Xu Z & Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 35:W265-268.

8. Oh DH, *et al.* (2009) Loss of halophytism by interference with SOS1 expression. *Plant Physiol* 151:210-222.

9. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* 29:e45.

Supporting Figures

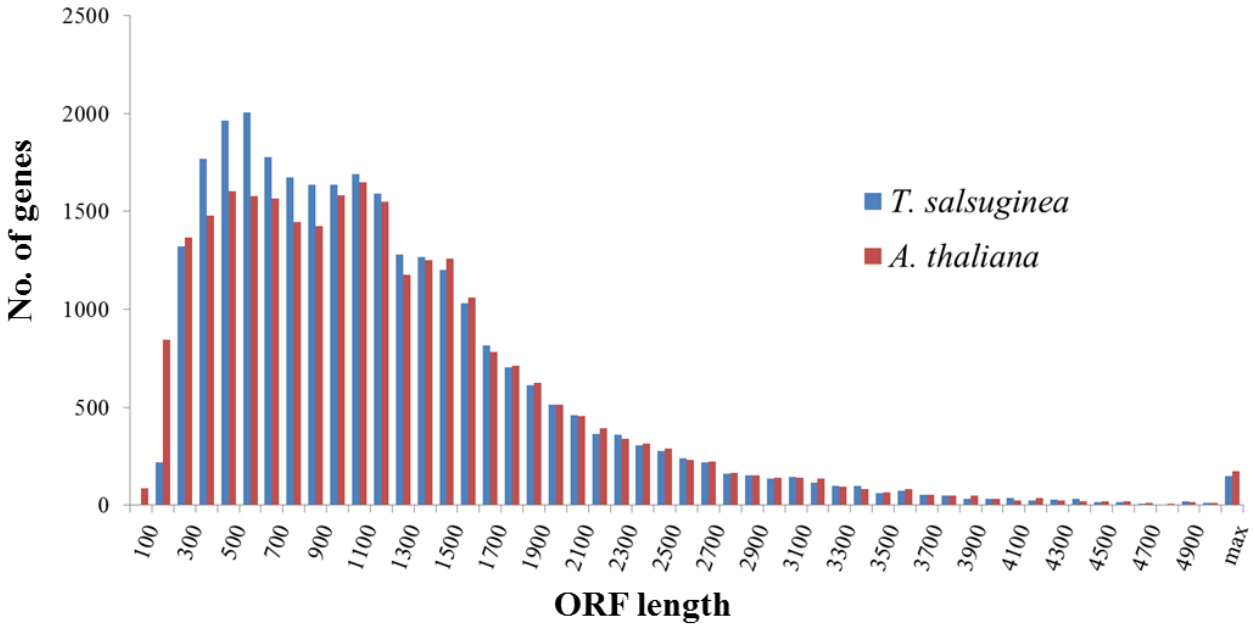


Fig. S1. ORF length distribution comparison between *T. salsuginea* and *A. thaliana*.

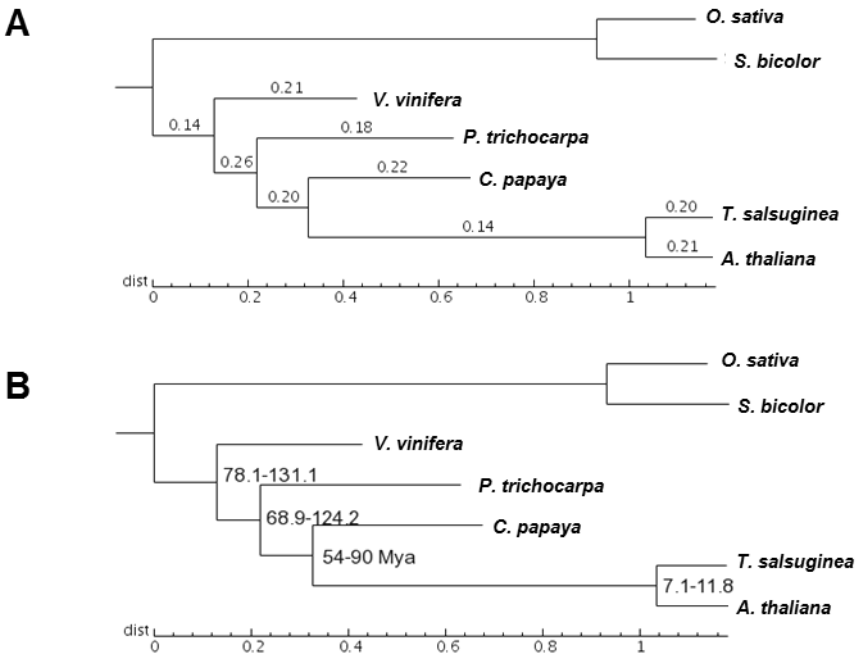


Fig. S2. Phylogenetic tree and estimation of species divergent time.

A. Phylogenetic tree of selected plant species constructed with 2226 single-copy gene families on 4-fold degenerate sites. The branch length represents the neutral divergence rate. Numbers shown on the branches represent the dN/dS rate of each branch. The posterior probabilities (credibility of the topology) for inner nodes are all 100%.

B. Estimation of divergent time. The numbers on the nodes identify the divergent time from the present (million years ago, Mya). The calibration time (fossil record time) interval (54-90 Mya) for Capparales was taken from published reports (Wikström, 2001; Crepet, 2004).

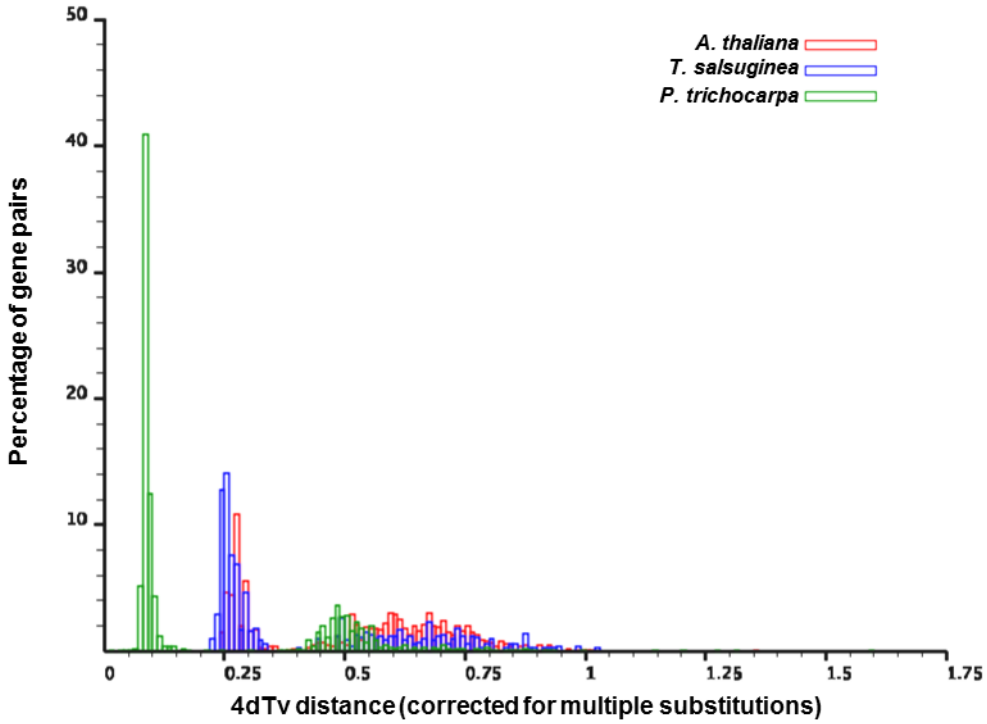
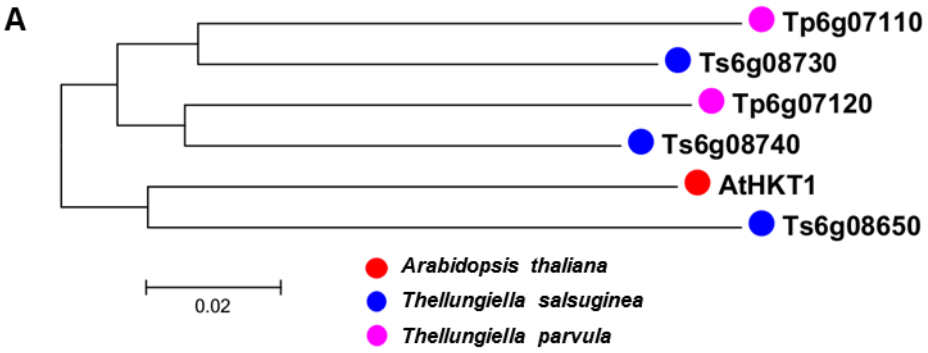
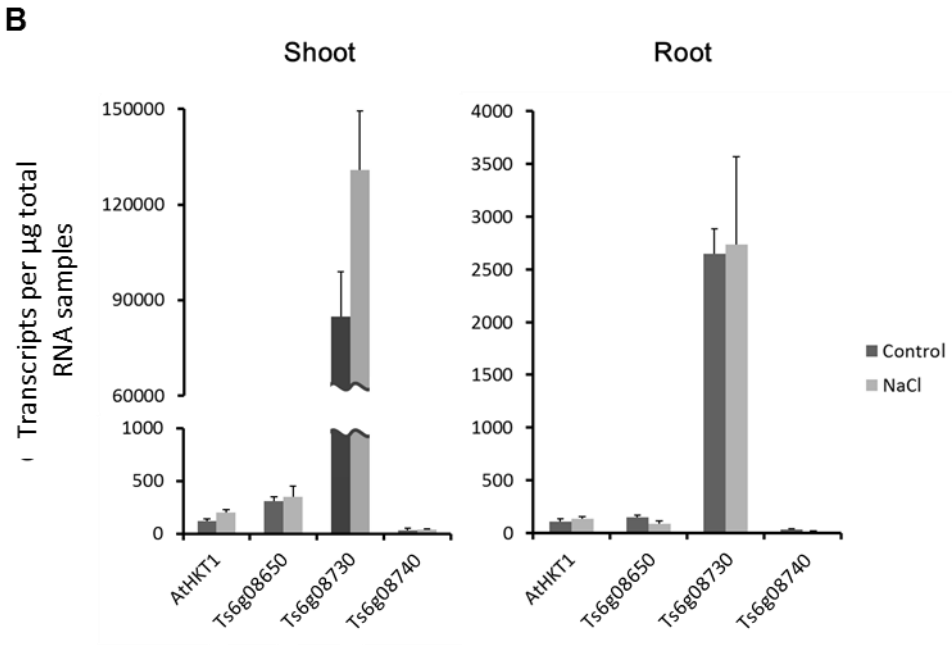


Fig. S3. 4dtv distance distribution for *T. salsuginea*, *A. thaliana* and *P. trichocarpa*.
The intra-genomic syntenic blocks among *T. salsuginea*, *A. thaliana*, and *P. trichocarpa* were detected using Mcscan program. The intervening gene number cutoffs in each block are 10 for *T. salsuginea* and *A. thaliana*, and 8 for *P. trichocarpa*, respectively. The 4dtv distances are calculated based on 4-fold degenerate sites following the HKY substitution model.


A
Tp6g07110
Ts6g08730
Tp6g07120
Ts6g08740
AtHKT1
Ts6g08650
0.02
Arabidopsis thaliana
Thellungiella salsuginea
Thellungiella parvula


B
Shoot
Root
Transcripts per µg total RNA samples
150000
120000
90000
60000
1000
500
0
4000
3500
3000
2500
2000
1500
1000
500
0
AtHKT1
Ts6g08650
Ts6g08730
Ts6g08740
Control
NaCl

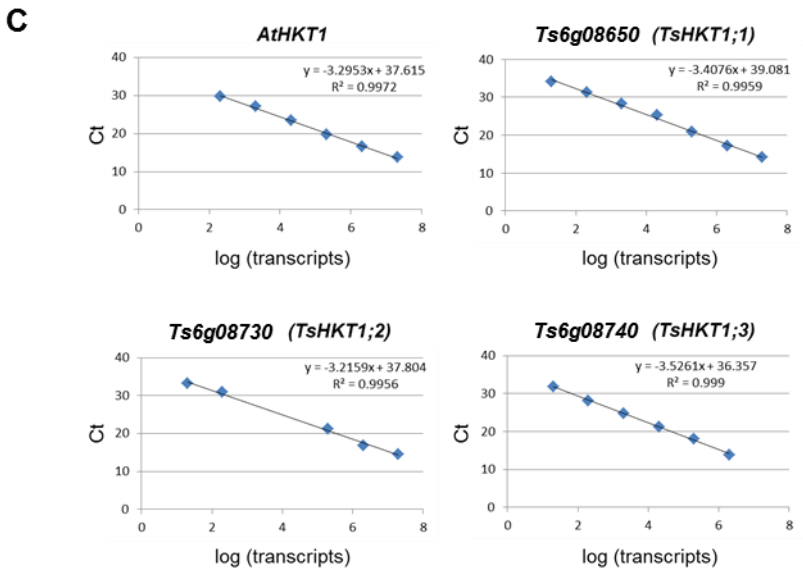

C
AtHKT1
y = -3.2953x + 37.615
R² = 0.9972
Ts6g08650 (TsHKT1;1)
y = -3.4076x + 39.081
R² = 0.9959
Ts6g08730 (TsHKT1;2)
y = -3.2159x + 37.804
R² = 0.9956
Ts6g08740 (TsHKT1;3)
y = -3.5261x + 36.357
R² = 0.999
Ct
log (transcripts)

Fig S4 Phylogenetic and expression analysis of *HKT1* genes.

A. Phylogenetic analysis of plant *HKT1* genes identifies three gene groups (Class I, II and III).

B. Quantification of transcripts of *HKT1* homologs from *A. thaliana* and *T. salsuginea*. RNA samples from 2 week-old *A. thaliana* and 3 week-old *T. salsuginea* plants treated with 200 mM NaCl for 12 hours were subjected to quantitative real-time RT-PCR as described in *SI Appendix*.

C. Standard calibration curves used for deducing the absolute transcript copy numbers from the real-time RT-PCR results. For detailed methods, see the *SI Appendix* and references therein.

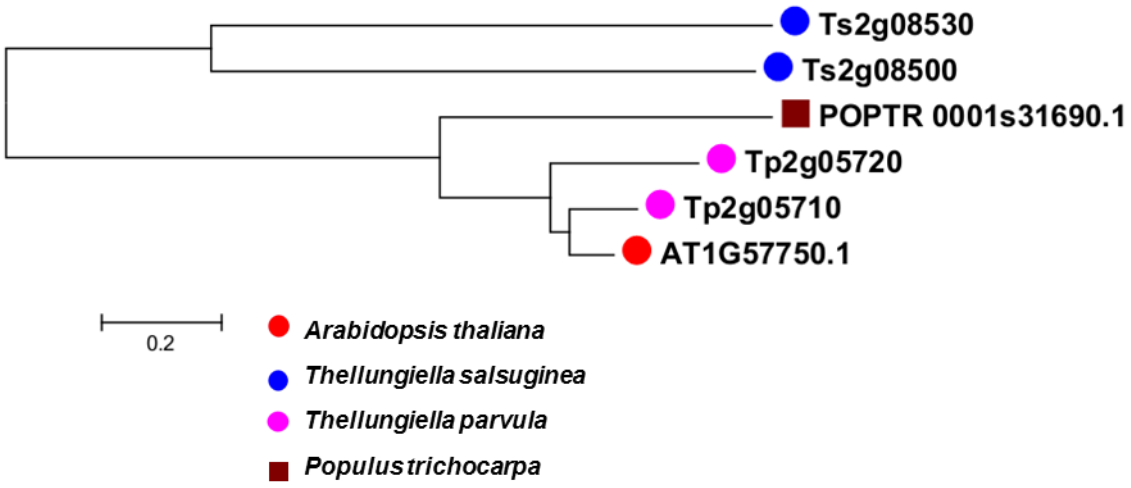


Fig. S5. Phylogenetic analysis of *MAH1/CYP96A15* genes in *T. salsuginea, A. thaliana, T. parvula, P. trichocarpa.*

The phylogenetic tree was constructed using the Neighbor Joining Method with the Mega 5.0 software. The MAH1/CYP96A15 gene, which belongs to the P450 gene family and functions as a key enzyme in the alkane-forming pathway, is tandem duplicated in both *T. salsuginea* and *T. parvula*. We failed to find the corresponding MAH1 genes in *V. vinifera*, *C. papaya and O. sativa*.

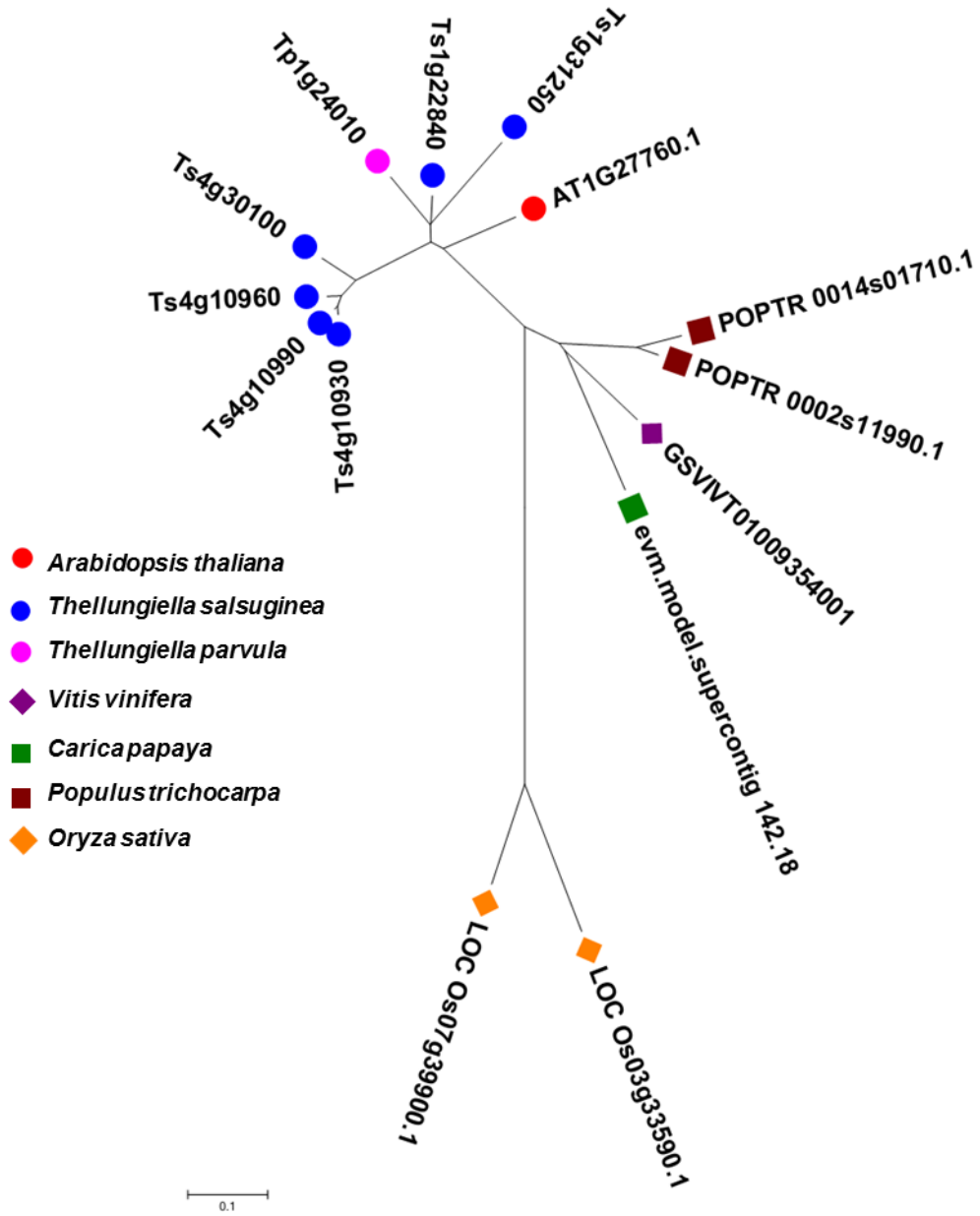


Fig. S6. Phylogenetic analysis of *SAT32* genes in *T. salsuginea, A.thaliana, T. parvula, V. vinifera, P. trichocarpa, C. papaya, O. sativa*.

The phylogenetic tree was constructed using the Neighbor Joining Method with the Mega 5.0 software.

Supporting Tables

Table S1. Features of the *T. salsuginea* genome.

Feature	Value
Estimated genome size	260 Mb
Assembled genome sequence	233,653,061 bp
Length of scaffolds in seven chromosomes	186,126,548 bp
Number of scaffolds anchored to chromosomes	515
Number of unplaced scaffolds	2167
Length of unplaced scaffolds	47,526,513 bp
Total number of scaffolds	2682
N50	403,516 bp
Number of scaffolds at least N50	119
Transposable elements (percentage)	121,046,173 bp (51.81%)
DNA transposons	20,160,164 bp (8.63%)
Retrotransposon	90,570,024 bp (38.76%)
Other	10,315,985 bp (4.42%)
Number of genes	28,457
Length of coding regions (percentage)	58,138,525 bp (24.88%)
Gene density	122 genes per Mb
Average gene length	2,041 bp
Average protein length	398 aa
Number of exons (per gene)	149,079 (5.23)
Average exon length	228 bp
Average intron length	200 bp
Gene annotation (percentage)	
InterPro	19,920 (69.9%)
GO	21,859 (76.8%)
With NCBI NR blast hit	26,016 (91.3%)
With ATH blast hit	25,288 (88.8%)
Unannotated	1,836 (6.45%)
Non-coding RNAs	
miRNA	162
tRNA	447
rRNA	11
snRNA	432

Table S2. Summary of the *T. salsuginea* genome sequencing data. The estimated genome size of 260 Mb is used to calculate the sequencing depth.

Insert size (bp)	Average read size (bp)	No. of sequencing lanes	No. of usable reads (Million)	No. of usable bases (Mb)	Sequencing depth (fold)
180	90	1	47.90	4311.56	16.58
200	41	3	98.75	4048.91	15.57
340	60	1	51.01	3060.76	11.77
374	75	2	84.65	6349.08	24.42
682	75	2	65.25	4893.7	18.82
2000	44	3	93.48	4113.16	15.82
2000	44	1	20.19	888.21	3.42
5000	44	1	25.12	1105.31	4.25
5000	44	3	90.76	3993.57	15.36
10000	44	2	14.07	619.24	2.38
10000	44	1	16.83	740.55	2.85
10000	44	1	15.70	690.76	2.66
Total		21	623.73	34814.81	133.90

Table S3. Statistics of repeat sequences in the *T. salsuginea* genome.

Identification method	Type of repeats	On seven chromosomes	Unanchored	All
RepBase	Retroposon	19,017,116	9,762,654	28,779,770
	DNA transposon	4,721,481	1,065,990	5,787,471
	Other	230,210	32,886	263,096
TEdenovo	Retroposon	36,353,568	25,436,686	61,790,254
	DNA transposon	11,721,545	2,651,148	14,372,693
	Other	7,256,637	2,796,252	10,052,889
Total repeats		79,300,557 (43%)	41,745,616 (88%)	121,046,173 (52%)

Table S4. Non-coding RNA genes in the assembled genome.

Type		Copy number	Average length(bp)	Total length(bp)
tRNA		447	74	33,154
rRNA		11	508	5,588
snRNA	CD-box snoRNA	323	99	31,919
	HACA-box snoRNA	37	124	4,589
	splicing	72	141	10,163
miRNA	Conversed	126	152	19,111
	Novel	36	118	4,252

Table S5. Functional comparison on different types of duplicated genes between *T. salsuginea* and *A. thaliana*. Blast2GO results of protein coding regions from *T. salsuginea* and *A. thaliana* were mapped to categories in the second level of GO terms. Fisher's exact test was performed to identify the significantly differed GO terms. P-values less than 0.05 and 0.01 are shown with light and dark grey circles, respectively. TD: tandem duplicated genes; SD: segmental duplicated genes; LTR: LTR retrotransposon carrying genes; RETRO: retrogenes.

Gene category	Total (ATH/TSA)	TD (ATH/TSA)	SD (ATH/TSA)	LTR (ATH/TSA)	RETRO (ATH/TSA)
biological regulation	5655/5368	480/495	2373/2253	65/87	74/96
carbon utilization	84/91	4/4	38/41	0/2	2/1
cell killing	11/17	6/9	2/2	0/1	0/0
cell proliferation	41/49	0/4	19/18	0/1	1/0
cellular component organization or biogenesis	1999/2248	154/202	819/893	21/34	27/32
cellular process	10594/11452	1125/1234	3885/4093	105/196	149/196
death	187/215	22/31	51/51	3/3	1/4
developmental process	2258/2724	175/227	987/1149	16/28	20/32
establishment of localization	2278/2394	224/237	898/921	15/21	18/25
growth	406/470	36/45	212/243	0/5	8/8
immune system process	357/398	55/62	132/107	10/11	2/5
localization	2364/2490	231/251	935/960	16/21	18/27
locomotion	17/28	3/6	6/13	2/0	0/0
metabolic process	9670/10308	1308/1269	3230/3436	92/190	138/175
multi-organism process	1135/1274	193/188	454/484	24/21	9/12
multicellular organismal process	2186/2677	175/242	926/1094	18/31	17/33
negative regulation of biological process	431/522	18/39	198/191	3/7	5/13
pigmentation	7/7	0/0	5/5	0/0	0/0
positive regulation of biological process	461/480	32/42	213/200	4/13	3/8
regulation of biological process	4778/5065	386/464	2034/2130	55/84	64/95
reproduction	1230/1490	91/155	497/577	8/15	8/16
reproductive process	1202/1463	88/149	489/567	7/15	8/16
response to stimulus	5412/6049	737/813	2208/2417	63/91	93/134
rhythmic process	61/101	0/6	39/56	0/0	0/0
signaling	1713/1737	201/208	694/747	33/32	36/38
viral reproduction	11/22	3/0	5/11	2/0	0/0
cell	17451/16118	1960/1624	6020/5822	232/242	252/258
cell junction	24/28	0/0	16/16	0/0	0/0
extracellular region	654/637	141/108	247/277	7/8	14/14
extracellular region part	53/73	9/5	25/43	1/0	0/1
macromolecular complex	4393/4382	278/312	1900/1826	41/47	64/70
membrane-enclosed lumen	2579/2556	175/191	1239/1160	27/33	36/36
organelle	10496/10750	833/931	3823/3837	104/160	123/172
symplast	17/18	0/0	11/10	0/0	0/0
virion	15/3	0/0	4/3	0/0	0/0
antioxidant activity	162/155	28/20	55/56	1/1	1/1
binding	12951/13785	1359/1478	4396/4718	130/312	171/246
catalytic activity	9228/9703	1273/1289	3052/3221	95/193	108/173
channel regulator activity	7/6	2/4	0/0	0/0	0/0
electron carrier activity	547/535	146/124	175/156	4/17	8/24
enzyme regulator activity	380/375	28/35	171/157	3/7	1/1
metallochaperone activity	4/3	0/0	2/0	0/0	0/0
molecular transducer activity	419/429	50/52	163/161	3/5	3/3
nucleic acid binding transcription factor activity	1734/1669	108/118	871/840	14/19	17/22
nutrient reservoir activity	67/56	28/20	19/24	3/1	0/0
protein binding transcription factor activity	61/66	2/4	18/22	0/0	1/1
protein tag	5/4	1/2	2/3	0/0	0/0
receptor activity	294/344	30/47	116/136	3/1	3/2
structural molecule activity	567/545	43/35	245/222	9/5	2/8
translation regulator activity	3/3	0/0	0/0	0/0	0/0
transporter activity	1347/1405	167/167	512/524	9/11	11/22
Total genes	**27416/28457**	**2708/2723**	**8429/8178**	**444/842**	**353/535**

Table S6. Comparison of transcription factor gene families between *T. salsuginea*, *T. parvula* and *A. thaliana*.

Gene Family	No. of genes		
	T. salsuginea	***T. parvula***	***A. thaliana***
RAV	9	6	6
NF-X1	3	2	2
EIL	9	8	6
LSD	4	4	3
ARR-B	18	21	14
G2-like	53	47	42
Nin-like	17	14	14
GRAS	40	35	33
HSF	28	23	24
CAMTA	7	6	6
E2F/DP	9	9	8
CPP	9	6	8
GRF	10	10	9
AP2	20	16	18
B3	69	59	64
Trihelix	31	29	29
M-type	70	52	66
MIKC	44	42	42
GATA	31	31	30
HD-ZIP	49	55	48
bZIP	75	76	74

Note: the TF data were downloaded from: http://planttfdb.cbi.pku.edu.cn/index.php?sp=At.

RAV Family: RAV transcription factor were strongly induced after pathogen infection and salt (PMID: 16927203) & RAV transcription factor were induced by cold stress (PMID: 15728337).

NF-X1 Family: The AtNFXL1 gene encodes a NF-X1 type zinc finger protein required for growth under salt stress (PMID: 16905136).

GRAS Family: involves in plant development regulation. RGL3 transcript levels were transiently increased by cold (PMID: 18757556).

HSF Family: heat stress factors. Salt and osmotic stress induced *HsfA2* gene expression, and *HSFA2* overexpression mutant showed enhanced osmotic stress (PMID: 17890230).

Trihelix Family: The transcript level of *OsGTγ-1* was strongly induced by salt stress, and overexpression of *OsGT γ-1* in rice enhanced salt tolerance at the seedling stage (PMID: 20039179).

EIL : ethylene. **LSD:** PCD. **ARR-B:** cytokinin. **G2-like:** chloroplast development. **Nin-like:** root nodules. **CAMTA:** calmodulin binding TF. **E2F/DP:** cell proliferation. **CPP:** cell division. **GRF:** growth regulation. **AP2:** development. **B3:** includes LAV, REM and RAV family. M-type&MIKC: MADS-box TFs. **GATA:** light responsive. **HD-ZIP:** development.

Table S7. Species distribution analysis of ionic homeostasis related gene families.

Gene Family	No. of genes		
	T. salsuginea	***T. parvula***	***A. thaliana***
NHX	8	11	8
HKT1	3	2	1
Shaker	9	9	9
KEA	6	6	6
KUP-HAK-KT	13	18	13
CNGC	27	21	20
TPK	4	7	6
PPa	7	6	6
AHA	10	10	11
ACA	16	12	11
ECA	3	4	4
CHX	28	28	29
CAX	5	5	6
AVP	4	3	2
VHA.a	3	3	3
VHA.c’	4	4	5
VHA.c’’	1	2	2
VHA.d	2	2	2
VHA.e	1	2	2
VHA-A	1	1	1
VHA-B	3	4	3
VHA-C	1	1	1
VHA-D	1	1	1
VHA-E	3	4	3
VHA-F	1	1	1
VHA-G	3	3	3
VHA-H	1	1	1
GLR	12	14	20
CCC	1	1	1
ATBGL	49	39	46
CBL	9	10	10
CIPK	30	28	25
CDPK	37	36	34

Table S8. Species distribution analysis of wax biosynthesis gene families.

Gene Family	No. of genes		
	T. salsuginea	***T. parvula***	***A. thaliana***
ACC	4	2	2
FATB	1	1	1
LACS	11	9	9
KCS	22	24	21
KCR	3	2	2
HCD	1	1	1
ECR	1	1	1
FAR	9	10	8
WS/DGAT	11	16	11
MAH1/CYP96A15	2	2	1
WBC11	1	1	1
CER5/WBC12	1	2	1
CER1/CER-like	3	3	4
CER2	1	0	1
CER3/WAX2/YRE/FLP1	1	1	1
CER7	1	1	1
WIN1/SHN1	1	0	1
Total	74	76	67

Table S9. Species distribution analysis of ABA biosynthesis and ABA signaling related gene families.

Gene Family	No. of genes		
	T. salsuginea	***T. parvula***	***A. thaliana***
ZEP	2	1	1
AAO	7	4	4
ABA3	1	1	1
NCED	7	7	7
CYP707A	5	4	4
SDIR1	1	1	1
PP2C	75	74	74
SNRK2	9	11	10
ABF	4	4	4
ABI5	1	1	1
AFP	4	4	4

Table S10. Species distribution analysis of other gene families related to salinity, drought and cold stress response or tolerance.

Gene Family	No. of genes		
	T. salsuginea	***T. parvula***	***A. thaliana***
PLD	15	11	12
P5CDH	1	1	1
P5CS	2	2	2
PDH	2	2	2
DREB	56	55	56
ERF	59	67	62
MAPK	18	19	20
MAPKK	10	11	10
MEKK	20	20	21
ZIK	11	11	11
Raf	45	50	48
AHK1	1	2	1
SKB1	3	2	1
SIZ1	2	2	1
LEA	42	41	40
OTS	2	3	2
ATSAT32	6	1	1