Genomics 94 (2009) 196–203

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Comparative sequence analysis of the *SALT OVERLY SENSITIVE1* orthologous region in *Thellungiella halophila* and *Arabidopsis thaliana*☆,☆☆

Gyoungju Nah [a,b,1], Christopher L. Pagliarulo [a,2], Peter G. Mohr [a,3], Meizhong Luo [b,4], Nick Sisneros [b], Yeisoo Yu [b], Kristi Collura [b], Jennifer Currie [b], Jose Luis Goicoechea [a,b], Rod A. Wing [a,b,*], Karen S. Schumaker [a,*]

[a] *Department of Plant Sciences, University of Arizona, Tucson, AZ 85721-0036, USA*
[b] *Arizona Genomics Institute, University of Arizona, Tucson, AZ 85721-0036, USA*

ARTICLE INFO

Article history:
Received 25 October 2008
Accepted 19 May 2009
Available online 28 May 2009

Keywords:
Thellungiella halophila
SALT OVERLY SENSITIVE1
Salt tolerance
Comparative sequence analysis
Microcolinearity
Simple sequence repeats

ABSTRACT

To provide a framework for studies to understand the contribution of SALT OVERLY SENSITIVE1 (SOS1) to salt tolerance in *Thellungiella halophila*, we sequenced and annotated a 193-kb *T. halophila* BAC containing a putative *SOS1* locus (*ThSOS1*) and compared the sequence to the orthologous 146-kb region of the genome of its salt-sensitive relative, *Arabidopsis thaliana*. Overall, the two sequences were colinear, but three major expansion/contraction regions in *T. halophila* were found to contain five Long Terminal Repeat retrotransposons, MuDR DNA transposons and intergenic sequences that contribute to the 47.8-kb size variation in this region of the genome. Twenty-seven genes were annotated in the *T. halophila* BAC including the putative *ThSOS1* locus. *ThSOS1* shares gene structure and sequence with *A. thaliana SOS1* including 11 predicted transmembrane domains and a cyclic nucleotide-binding domain; however, different patterns of Simple Sequence Repeats were found within a 540-bp region upstream of *SOS1* in the two species.

Introduction

Excess salt in the soil affects the growth of most plants several ways. Uptake of sodium ions alters the balance of sodium and potassium in cells and reduces potassium-based metabolic activity [1]. The build-up of salt in the soil also alters the soil water potential making it difficult for the plant to continue to take up water required for sustained growth [1]. *Thellungiella halophila* (*T. halophila*), a member of the *Brassicaceae*, has recently emerged as a model for understanding plant adaptation to growth in saline conditions. As a halophyte, *T. halophila* is able to grow in high salt concentrations, conditions that inhibit the growth of its salt-sensitive relative *Arabidopsis thaliana* (*A. thaliana*, a glycophyte) as well as the growth of most crop plants [2–3]. *T. halophila* has a small diploid genome (240 Mb ($2n=14$)) [3], a short and self-fertile life cycle and an ability to be transformed by floral dipping with *Agrobacterium tumefaciens*. Evolutionary studies have shown that *A. thaliana* and *Brassica* split from a common ancestor approximately 14.5 to 20.4 MYA [4]. *T. halophila* forms a clade separate from other members of the *Brassicaceae*, implying that it originated from a different lineage after divergence from a common ancestor of *A. thaliana* [5–8] and suggesting that salt tolerance originated from a distinct lineage.

Membrane transport proteins have been shown to be important components of salt tolerance mechanisms due to their regulation of ion homeostasis. A well-defined pathway for regulation of sodium ion homeostasis during plant growth in salt in *A. thaliana* is the SALT OVERLY SENSITIVE (SOS) pathway [2,9,10]. In this pathway, a calcium-binding protein, SOS3, perceives a change in intracellular calcium induced by salt stress and then binds to and activates SOS2, a serine-threonine protein kinase. The SOS3–SOS2 complex increases the expression and the activity of SOS1, a plasma membrane Na^+/H^+ exchanger (antiporter) [11,12]. Activated SOS1 transports cytosolic sodium out of the cell, reducing the cellular build-up of toxic levels of sodium [10]. Recent studies have provided insight into the biochemical changes that take place when *T. halophila* is grown in salt [13]. Sodium accumulation appears to be regulated with highest levels in old leaves, followed by young leaves and taproots and lowest levels in lateral roots. The H^+ transport and hydrolytic activities of the vacuolar

☆ This research was supported by grants from the Bud Antle Endowed Chair to R.A.W. and the United States Department of Agriculture CSREES/Southwest Consortium on Plant Genetics and Water Resources (06N08/07R08) and the Department of Energy (DE-FG02-04ER15616) to K.S.S.

☆☆ DNA sequence in this article has been deposited to the NCBI database with Accession Number FJ386403.

* Corresponding authors.
E-mail addresses: rwing@ag.arizona.edu (R.A. Wing), schumake@ag.arizona.edu (K.S. Schumaker).

[1] Present address: Section of Molecular Cell and Developmental Biology, The University of Texas at Austin, One University Station, A-4800, Austin, TX 78712-0159, USA.

[2] Present address: Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721-0036, USA.

[3] Present address: Australian Animal Health Laboratory, CSIRO Livestock Industries, PO Bag 24, Geelong VIC 3220, Australia.

[4] Present address: National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China.

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.05.007

(tonoplast) and plasma membrane H^+-ATPases and the Na^+/H^+ antiport activity at the tonoplast are enhanced in salt-grown *T. halophila*, suggesting a link between salt tolerance and regulation of sodium ion homeostasis. Recently, a *T. halophila SOS1* cDNA sequence was compared to *SOS1* sequences from *A. thaliana* and the halophyte *Mesembryanthmum crystallinum* [14]. Amino acid sequence comparisons indicated that the *SOS1* coding regions are relatively well-conserved with all sequences containing the typical SOS1 transmembrane and cyclic nucleotide-binding domains. To functionally link ThSOS1 to salt tolerance, *ThSOS1* knock down lines were generated using an RNAi strategy [14]. When the growth of RNAi lines with greater than a 50% reduction in *ThSOS1* transcript accumulation were compared to the growth of wild-type *T. halophila*, mature RNAi plants showed no difference in growth under control (no salt) conditions. However, when grown in the presence of 350 mM NaCl, the RNAi lines exhibited severe salt-sensitivity and resembled *A. thaliana* grown in salt [14]. This result indicates that, as in *A. thaliana*, *SOS1* in *T. halophila* plays a significant role in its ability to grow in salt and suggests that altered levels of *SOS1* expression might be important for differences in glycophyte/halophyte growth in response to salt [14]. Evidence for the importance of the temporal and spatial expression of SOS1 in glycophyte/halophyte growth differences comes from analysis of *ThSOS1* expression in the presence and absence of salt [15] and from comparative studies in which shoot *ThSOS1* expression was shown to be more strongly induced under salt stress and root *ThSOS1* expression constitutively higher in non stress conditions when compared to *AtSOS1* expression in *A. thaliana* [16].

Comparative genomics has been used to identify biologically significant regions of genomes, for positional cloning of genes and for studies of the evolutionary history of whole genomes or genomic regions [17–20]. Comparative analyses between *A. thaliana* and related species like *T. halophila* with novel traits should provide important information about recent genome diversification, genetic variation within the *Brassicaceae* and ultimately insight into the origin of this variability.

To provide a framework for studies to link *SOS1* expression to differential salt tolerance, we carried out a comparative genomic analysis to determine if differences in the non-coding region of *SOS1* are found in the two species. A 193 kb *T. halophila* BAC clone containing the putative *SOS1* locus was sequenced, annotated and compared with sequence in the orthologous 146 kb region of the *A. thaliana* genome on chromosome 2. This comparative sequence analysis provided insight into the structure and organization of the *T. halophila* genome, determined the complete structure of the *T. halophila SOS1* gene and identified putative *T. halophila SOS1*-specific genomic features.

Results

Overall sequence comparison of the SOS1 orthologous region in T. halophila and A. thaliana

To investigate the genome structure of the *SOS1* locus and its surrounding region in *T. halophila*, we sequenced BAC ThSBa0001B18 (NCBI GenBank Accession No. FJ386403; 193,021 bp) identified by hybridization of a *ThSOS1* probe to a *T. halophila* BAC library. The orthologous region on chromosome 2 of *A. thaliana* was 146,312 bp in size. Unless indicated, the comparisons we report refer to these genomic regions from the two species. The overall GC content was 35% in *T. halophila* and 33% in *A. thaliana*, similar to the overall GC content of 35.5% reported for chromosome 2 of *A. thaliana* [21]. To examine the overall colinearity, the *T. halophila* BAC sequence was aligned to the orthologous *A. thaliana* sequence using a dotplot alignment program. As shown in Fig. 1, a high level of conservation was observed between the two species; however, several local genome rearrangements, which disrupt the linear pattern, were detected. Most of these disruptions were due to expansions/contractions, duplications and inversions (Fig. 1, Supplemental Table 1).

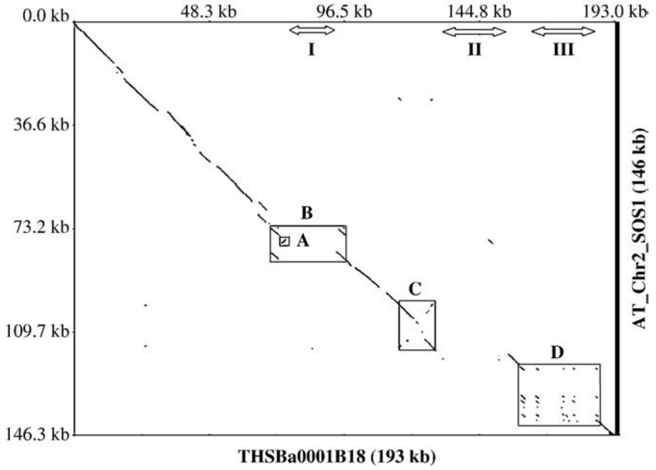

Fig. 1. Dotplot alignment between *T. halophila* BAC (FJ386403) and *A. thaliana* (AT_Chr2_SOS1). Expansion/contraction regions with over 5 kb difference are indicated as I–III. Non-colinear regions with genomic rearrangements that include coding sequences are designated as A–D.

Gene prediction and gene structure

A total of 27 Open Reading Frames (ORFs) from *T. halophila* and 30 ORFs from *A. thaliana* were predicted and grouped into 20 orthologous gene sets based on gene prediction software, full-length cDNAs, Expressed Sequence Tags (ESTs) and homology searches against the *A. thaliana* protein database (Fig. 2, Supplemental Table 2). Predicted genes were defined from the translational start codon to the stop codon. All genes or their encoded proteins showed *E*-values less than E^{-17} based on sequence similarity analyses of *A. thaliana* ESTs and proteins using TBLASTN and BLASTP, respectively.

The average gene density was 1 gene per 7.1 kb for *T. halophila* and 1 per 4.9 kb for *A. thaliana* (Supplemental Table 3). The gene density calculated for the 26,819 genes in the whole *A. thaliana* genome was 1 gene per 4.4 kb [22]; this 0.5 kb difference indicates that average gene density around the *SOS1* locus is lower than the average gene density at whole genome level. To estimate the gene number for the *T. halophila* genome, its genome size was divided by the *T. halophila* gene density, resulting in a gene count of 33,576 genes, approximately 6800 more genes than in *A. thaliana*.

The average gene size was 2321 bp for *T. halophila* and 2073 bp for *A. thaliana* while the average peptide size was 442 amino acids (aa) in *T. halophila* and 420 aa in *A. thaliana*. Data for the whole *A. thaliana* genome indicated an average gene size of 2221 bp and an average predicted protein length of 517 aa (TAIR 7 release) [22], indicating that estimated gene and protein sizes around the *SOS1* locus are smaller than the estimates at the whole genome level. When compared to the whole *A. thaliana* genome, *T. halophila* genes in this region have larger average gene sizes, but smaller average protein sizes. In the *SOS1* orthologous region, the average exon number was 5.3 per gene in *T. halophila* and 5 per gene in *A. thaliana*. The average exon size was 251 bp in both species, which is smaller than the average exon size of 268 bp calculated using the whole *A. thaliana* genome [22]. The average intron number was 4.3 per gene in *T. halophila* and 4 per gene in *A. thaliana*. Since the 5′ UTR and 3′ UTR regions of each gene were not defined in our annotation, only introns found within an open reading frame (ORF) were considered. The average intron size was 231 bp in *T. halophila* and 201 bp in *A. thaliana*. These values were greater than the average intron size of 165 bp calculated using the whole *A. thaliana* genome [22]. This comparison indicates that intron size is the primary determinant of gene size variation between *T. halophila* and *A. thaliana* in the *SOS1* orthologous region.

Transmembrane proteins play important roles in mediating communication in compartmentalized cellular environments. The

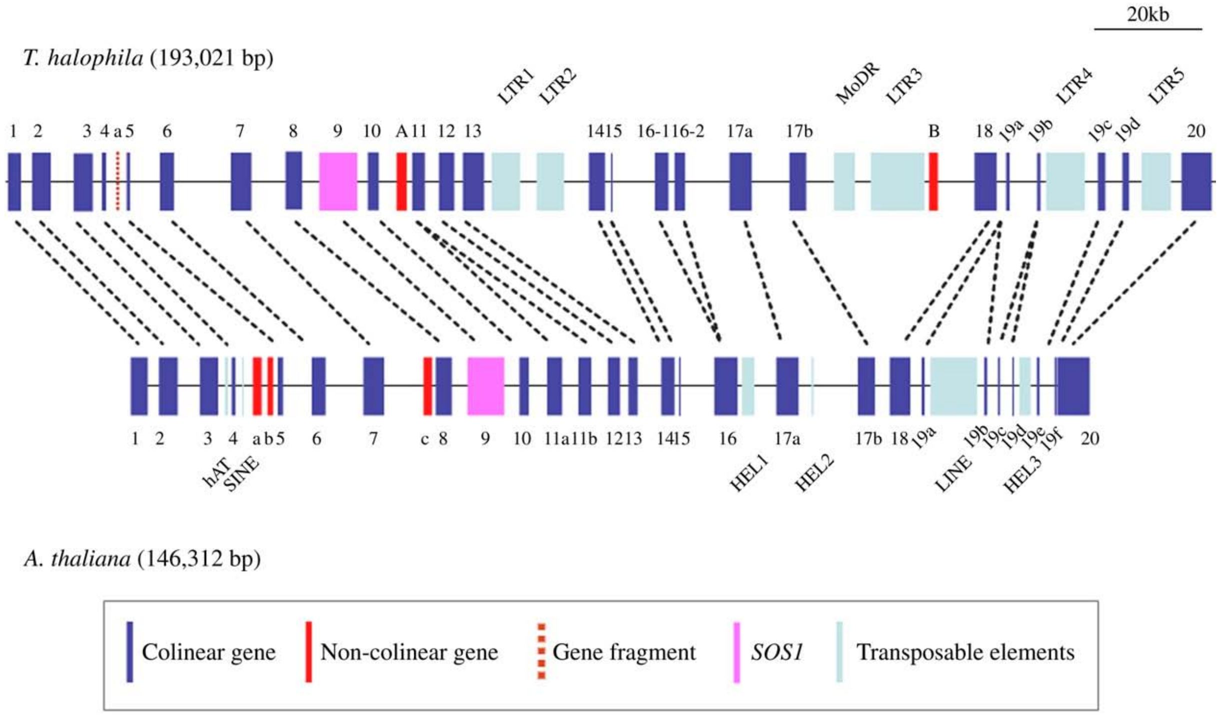
20kb
T. halophila (193,021 bp)
LTR1
LTR2
MuDR
LTR3
LTR4
LTR5
1 2 3 4 a 5 6 7 8 9 10 A 11 12 13 14 15 16-1 16-2 17a 17b B 18 19a 19b 19c 19d 20
1 2 3 4 a b 5 6 7 c 8 9 10 11a 11b 12 13 14 15 16 17a 17b 18 19a 19b 19c 19d 19e 19f 20
hAT
SINE
HEL1
HEL2
LINE
HEL3
A. thaliana (146,312 bp)
Colinear gene
Non-colinear gene
Gene fragment
SOS1
Transposable elements

Fig. 2. Genomic organization around the *SOS1* orthologous region in *T. halophila* and *A. thaliana*. Genes 11a–b, 17a–b and 19a–f represent tandemly duplicated genes. In *A. thaliana*, gene 16 was annotated as a single ORF with supporting EST evidence (TC282705) but in *T. halophila*, 16-1 and 16-2 were annotated as two separate ORFs because no gene model or EST evidence supported the fusion of two ORFs. LTR, LTR retrotransposon; SINE, Short Interspersed Transposable Element; LINE, Long Interspersed Transposable Element; MuDR, Mutator-like DNA transposon; Hel, Helitron; hAT, hAT DNA transposon; SSR, Simple Sequence Repeat.

presence of multiple genes with transmembrane domains in the *SOS1* region was predicted using the Simple Modular Architecture Research Tool (SMART). In this region of *T. halophila*, eight genes, Th07, Th08, Th09, Th12, Th14, and Th19a-Th19c were shown to contain transmembrane domains. In the orthologous region of *A. thaliana*, At07 (AT2G01950), At_c (AT2G01960), At08 (AT2G01970), At09 (*SOS1*, AT2G01980), At12 (AT2G02020), At14 (AT2G02040) and At19a–At19f (AT2G02100–AT2G02147) were shown to have transmembrane domains. Whether evolutionary forces maintained transmembrane gene clusters in this region remains to be determined.

Identification of transposable elements

Both Long Terminal Repeat (LTR) retrotransposable and DNA transposable elements were identified in the *SOS1* region of the two species (Table 1). Five intact LTR retrotransposable elements were identified in *T. halophila* based on both homology searches against the *A. thaliana* repeat database and structural identification (including homology between LTR pairs, the presence of 5′ TG—CA3′ in the beginning and end of an LTR, identification of 5 bp tandem duplication insertion sites and internal coding sequences). LTRs 1, 2, 4 and 5 (Fig. 2) were annotated as copia-type retrotransposons of 4754 bp, 4507 bp, 6429 bp and 4808 bp in size, respectively. LTR3 was annotated as 8839 bp of gypsy-type LTR retrotransposon. All five intact LTR retrotransposons were detected in intergenic regions and contributed to the disruption of microcolinearity. A 4622 bp-Mutator DNA transposon (MuDR)-like structure was found with 123 bp of Terminal Inverted Repeats (TIRs) and 9 bp of target site duplications (TTATTTTAT) flanking the TIRs. A Pfam program search showed that this MuDR-like structure contained a MuDR domain that encodes a transposase for Mutator transposable elements.

Table 1
Composition of transposable elements in the *T. halophila* BAC and its orthologous region in *A. thaliana*.

Transposable elements	*T. halophila*		*A. thaliana*		Location
	Number	Size (bp)	Number	Size (bp)	
I. Retrotransposon					
LTR retrotransposon					
Ty1/Copia type	4	20,498			Intergenic
Ty3/Gypsy type	1	8839			Intergenic
Non LTR retrotransposon					
LINE			1	7731	Intergenic
SINE			1	177	Intergenic
Percentage in the BAC		15.2%		5.4%	
II. DNA transposon					
MuDR	1	4622			Intergenic
Helitron			3	4280	Intergenic
hAT			1	531	Intergenic
Percentage in the BAC		2.4%		3.3%	
Total number	6		6		
Total size (bp)		33,959		12,719	
Total percentage		17.6%		8.7%	

In contrast, *A. thaliana* appears to have a completely different repertoire of repeat elements in the *SOS1* orthologous region. First, no intact LTR retrotransposons were identified while two types of non-LTR retrotransposons were identified: a 177 bp-SINE (Short Interspersed Nuclear Element)-like structure and a 7731 bp-LINE (Long Interspersed Nuclear Element)-like element. In addition, two types of DNA transposons were found: one hAT and three Helitrons. The hAT-like structure was confirmed by identification of 8 bp-target site duplications (TG/AAATACG). The three Helitrons were confirmed by detection of conserved termini of 5′-TC and CTAG-3′, their insertion at AT target sites and the presence of palindromic structure (an 18 bp-palindromic structure (TCCGCGGTATACCGCGGA) in the 11 bp upstream of the 3′ terminus of Helitron1, an 18 bp-palindromic structure (CCCGCGGTAAATTGCGGG) in the 11 bp upstream of the 3′ terminus of Helitron2 and a 17 bp-palindromic structure (CCTGCGG-TATACCGCGG) in the 12 bp upstream of the 3′ terminus of Helitron3).

The presence of five LTR retrotransposons and one MuDR in *T. halophila* suggests that these elements inserted in the present location after speciation. To determine if this is the case, we dated the insertion times of these LTR retrotransposons. At the time of insertion, the two LTRs of an intact LTR retrotransposon are assumed to be identical. Based on calculations of sequence divergence between two LTRs of an element, the insertion times can be estimated. The insertion times of LTR1, LTR3, LTR4 and LTR5 were 1.39 MYA, 0.47 MYA, 1.17 MYA and 1.14 MYA, respectively. The insertion time of LTR2 appeared to be most recent based on 100% sequence identity between a pair of LTRs and, therefore, could not be dated. All of the transposable elements identified in this analysis were located in intergenic regions in both *T. halophila* and *A. thaliana*. The total size of transposable elements was 33,959 bp in *T. halophila* and 12,719 bp in *A. thaliana*. This accounted for 17.6% of the region studied in *T. halophila* and 9% of the corresponding region in *A. thaliana*.

Simple sequence repeat analysis

Simple Sequence Repeat (SSR) composition was analyzed using the *A. thaliana* Simple Sequence Repeat Database and the SPUTNIK program. In the *SOS1* region of *T. halophila*, a total of 42 SSRs were found between and within genes, including four mononucleotides, 25 dinucleotides, 12 trinucleotides and one hexanucleotide (Supplemental Table 4). Dinucleotide SSRs were the major SSRs, and accounted for more than half of the total SSRs found in the *SOS1* region of *T. halophila*. Thirty out of 42 SSRs in *T. halophila* were found in intergenic regions, including two SSRs in LTR retrotransposons, six SSRs in exons and six SSRs in introns. In the orthologous region of *A. thaliana*, a total of 20 SSRs were detected, including one mononucleotide, 11 dinucleotides and eight trinucleotides. Dinucleotide SSRs were also the major SSRs in *A. thaliana*. Ten SSRs were found in intergenic regions, eight SSRs in exons and two SSRs in introns. Analysis of SSRs in intergenic regions uncovered two sets of SSRs with similar flanking sequences in orthologous positions. The TC/GA repeat in the intergenic region between gene 05 and gene 06 contained nine repeats in *T. halophila* and eight repeats in *A. thaliana*. The CAA/TTG repeat in the third exon of gene 17a (C2H2 zinc finger protein) contained 13 repeats in *T. halophila* and 15 in *A. thaliana*. SSRs identified in both species should serve as molecular genetic markers for future genetic studies.

Colinearity and disruptions

As shown in Fig. 2 and Supplemental Table 2, the *T. halophila* genomic region analyzed shared significant colinearity in terms of gene content, order and orientation relative to the orthologous region in *A. thaliana*. Of the 27 predicted ORFs from *T. halophila* and 30 from *A. thaliana*, six ORFs from *T. halophila* and ten from *A. thaliana* were present as duplicated copies. These duplicated copies were found in genes 11, 13, 17 and 19. All but one of the twenty sets of orthologous genes, gene 13, had identical orientations. To determine orthologous relationships, phylogenetic trees of duplicated genes 11, 17 and 19 were constructed (Supplemental Fig. 1).

Even though the two sequenced regions were highly conserved, there were several disruptions to this conservation, including the presence of transposable element insertions/deletions, non-colinear genes and gene duplications and inversions. Transposable element insertion in three regions was found to be the most significant cause of colinearity disruption and genome size variation between the two species (Fig. 1, Supplemental Table 1). Region I contained two copia-type LTR retrotransposable elements. Region II carried one gypsy-type LTR retrotransposable element and one MuDR DNA transposon, while Region III included two copia-type LTR retrotransposons. The total size of these three regions was 47.8 kb (Supplemental Table 1), which was similar to the 46.7 kb difference in the two orthologous regions.

Two non-colinear genes were identified in *T. halophila* (Th_A and Th_B) and three in *A. thaliana* (At_a, At_b and At_c). To explore the possibility that orthologous counterparts of these genes exist as partial genes that cannot be detected by gene prediction programs, TBLASTN was performed against the *A. thaliana* genome sequence. However, no traces of any partial genes were detected. A similar search was made with the *T. halophila* sequence, using the amino acid sequences of At_a, At_b and At_c via TBLASTN. In the *T. halophila* sequence, a gene fragment of At_a was found while gene fragments of At_b and the At_c were not detected. To investigate whether non-colinear genes in *T. halophila* are present in other locations in the *A. thaliana* genome, the *A. thaliana* protein database was searched using BLASTP. Two copies of Th_A (Valyl t-RNA synthetase) homologues were found on chromosome 1 in *A. thaliana* (AT1G14010, $1E^{-41}$ and AT1G27160, $8E^{-34}$) and Th_B (F-box protein) was found in multiple copies in the *A. thaliana* genome. The two genes with highest homology to Th_B were AT2G02030 ($2E^{-89}$) and AT2G05600 ($3E^{-69}$), suggesting that homologues of Th_A and Th_B exist in other locations in the *A. thaliana* genome. Since the *T. halophila* genome has not yet been sequenced, we were unable to determine whether homologues of non-colinear *A. thaliana* genes reside in other parts of the *T. halophila* genome.

Inversions and duplications were also shown to contribute to conservation disruption. Region A (1.3 kb) contained an inversion (Fig. 1, Supplemental Table 1) with a gene encoding a putative F-box protein in the sense orientation in *A. thaliana* but in the antisense orientation in *T. halophila* (Supplemental Table 2). Region B (25.9 kb) contained duplications (Fig. 1, Supplemental Table 1) with two copies of putative peptide transporter genes identified in both species. Region C (11.5 kb) contained duplicated genes that were inverted in both species; two copies of putative zinc finger protein genes were present in an inverted array (Fig. 1, Supplemental Table 1). Region D (27.6 kb) contained a tandem array of duplicated genes encoding putative protease inhibitors; the two species had different numbers of duplications in this region with four copies in *T. halophila* and six copies in *A. thaliana* (Fig. 1, Supplemental Table 1, Supplemental Table 2).

Evolutionary analysis of functional conservation and divergence time

To determine if functional conservation has been maintained between orthologous genes in these two related species, a Ka/Ks test was performed. Of the 20 sets of genes, those with sequence identity over 80% in both cDNA and amino acid composition were selected for further evolutionary analysis. Gene pairs with unclear orthology between the two species, such as the gene 19 duplicates, were excluded. A total of 17 orthologous gene sets were used in this analysis for Ka/Ks tests and estimation of divergence time of the two species. The Ka/Ks ratios for all 17 gene pairs were less than 1 (Table 2) indicating that functional conservation was maintained between these orthologous pairs. From our analysis, the approximate divergence time of *T. halophila* and *A. thaliana* was estimated to be between 10 and 14 MYA (Table 2) implying a more recent common ancestry for *T. halophila* and *A. thaliana* relative to *Brassica* and *A. thaliana* than has been reported previously [4]. Previous divergence time between *Brassica* and *A. thaliana* was calculated using information from mitochondrial gene sequences which are known to have silent nucleotide substitution rates that differ from nuclear-encoded gene sequences [23,24]. Therefore, it is likely that calculating divergence times using orthologous nuclear genes provides more precise information. To determine if selection was imposed on protein-coding sequences of paralogous genes, Ka/Ks tests among paralogous genes were performed for genes 11, 17 and 19. Their Ka/Ks ratios were also less than 1 (Supplemental Table 5), implying functional conservation between paralogous pairs.

Structure comparison of ThSOS1 and AtSOS1

Because the *SOS1* gene is a major determinant of salt tolerance in *A. thaliana*, extensive comparative analysis of the *SOS1* genes from *T. halophila* and *A. thaliana* was performed. First, to confirm orthology between ThSOS1 and AtSOS1, a neighbor joining phylogenetic tree

Table 2
Ratio of nonsynonymous (Ka) vs. synonymous (Ks) substitution rates in 17 orthologous genes and estimated divergence time.

T. halophila	*A. thaliana*	Putative function	Ka/Ks	Divergence time (MYA[a])
Th01	At01	Putative phosphatase	0.212	10.26
Th02	At02	Cyclin-like protein	0.201	12.28
Th03	At03	Putative microtubule-associated protein	0.1385	10.4
Th04	At04	Unknown	0.3554	9.44
Th06	At06	Putative C2H2-type zinc finger protein	0.156	11.87
Th07	At07	Brassinosteroid receptor-like protein	0.0713	15.11
Th08	At08	Putative endomembrane protein	0.0327	12.78
Th09	At09	Putative Na^+/H^+ antiporter	0.2857	10.42
Th10	At10	Unknown	0.3201	9.93
Th11	At11b	Putative glutamate decarboxylase	0.0559	13.75
Th12	At12	Putative peptide transporter	0.1039	14.96
Th14	At14	Putative peptide transporter	0.0378	14.13
Th15	At15	Putative NADH dehydrogenase	0.157	12.76
Th17a	At17a	Putative C2H2-type zinc finger protein	0.1474	7.43
Th17b	At17b	Putative C2H2-type zinc finger protein	0.1356	11.61
Th18	At18	Putative helicase	0.1147	10.06
Th20	At20	Putative PPR repeat protein	0.1753	13.07
Mean				11.78
SD[b]				2.11

[a] Million years ago.
[b] Standard deviation.

was constructed using eight members of the NHX family from *A. thaliana* and ThSOS1 with a bootstrap value of 1000 (Supplemental Fig. 2). ThSOS1 clustered with AtSOS1 (NHX7) and separated from all other *A. thaliana* NHX gene family members, indicating that ThSOS1 was orthologous to AtSOS1. The *ThSOS1* gene (from the ATG to the stop codon) was 6267 bp in length and encoded a predicted protein sequence of 1146 aa, while *AtSOS1* was 6076 bp long and encoded a predicted 1146 aa protein (Fig. 3). Both genes had the same total exon size of 3441 bp, but had different total intron sizes of 2826 bp for *ThSOS1* and 2635 bp for *AtSOS1*, indicating that the 191 bp gene size difference was due to differences in intron size. Both genes had 23 exons with an average size of 150 bp. Average intron sizes were 129 bp in *ThSOS1* and 120 bp in *AtSOS1*. Peptide sequence identity was 83% and cDNA sequence identity was 87%.

The three main domains that are important for *AtSOS1* function were also found in *ThSOS1*. According to Pfam analysis, *ThSOS1* contained the Na^+/H^+ exchange domain (aa 31 to 444), 11 transmembrane domains (Supplemental Fig. 3; aa 50 to 69, 86 to 105, 123 to 142, 155 to 174, 191 to 120, 136 to 155, 184 to 203, 221 to 239, 257 to 276, 293 to 312 and 325 to 343) and a cyclic nucleotide-binding domain (aa 758 to 843).

A significant structural difference between the two genes was the presence of three SSRs in the 5′ upstream region of *ThSOS1*. $(TCA)_8$, $(CTT)_{18}$ and $(TA)_{12}$ were identified within 540 bp upstream of the *ThSOS1* putative translational start site (Fig. 3). No unique SSRs were identified in the corresponding region in *AtSOS1*, but a $(CTT)_3$ repeat was detected in the *AtSOS1* 5′ UTR.

Discussion

Gene colinearity is generally conserved in the SOS1 orthologous region, but a few exceptions exist

Analysis of gene homology, order, orientation and physical distance for a single orthologous region of the *T. halophila* and *A. thaliana* genomes revealed that they share a high degree of colinearity. First, both regions contain 20 putative gene sets that shared a high level of sequence similarity ($<E^{-20}$) in both amino acid and cDNA sequence. Second, these 20 genes are in the same order (colinear) in the two species. Third, the transcriptional orientation of all 20 genes is the same except for gene 13 which has an opposite transcriptional orientation. Fourth, all 20 genes are organized within a similar

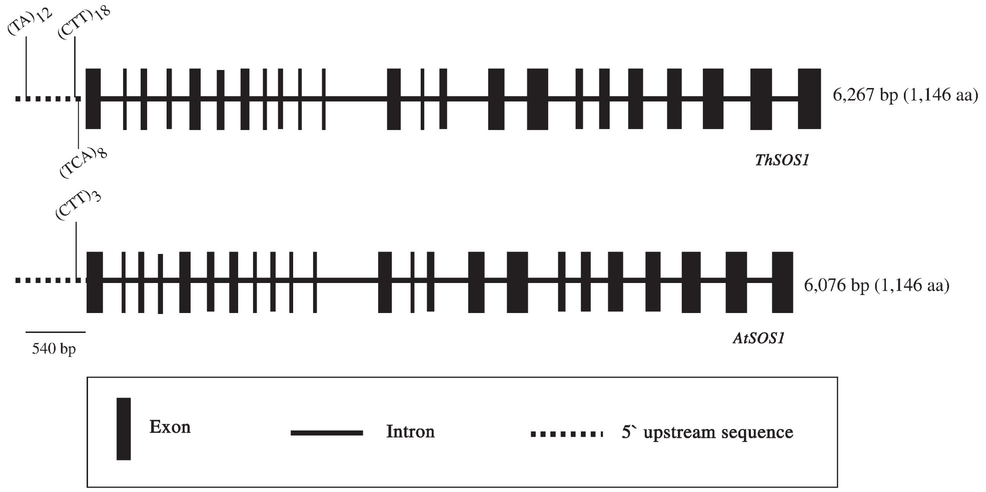

Fig. 3. Gene structure comparison of *ThSOS1* and *AtSOS1*. Exons are indicated as filled boxes and introns are represented as lines. Location of SSRs in the 5′ UTR of *ThSOS1* and *AtSOS1* are indicated.

physical space (146 kb) in both species (excluding the intergenic regions containing 5 LTR retrotransposable elements in *T. halophila*).

In spite of this conservation of gene colinearity, comparative analysis revealed the presence of two non-colinear genes in *T. halophila* and three in *A. thaliana*. For example, an orthologue of Th_A (Valyl t-RNA synthetase) in *T. halophila* is absent in the *SOS1* orthologous region in *A. thaliana* (chromosome 2), but is present in a paralogous region on *A. thaliana* chromosome 1. Another non-colinear gene in *T. halophila*, Th_B encodes an F-box protein and F-box homologues are present as many copies throughout the *A. thaliana* genome. In *A. thaliana*, At_a (AT2G01918), At_b (AT2G01920) and At_c (AT2G01960) were identified as non-colinear genes. While it is currently not possible to determine if orthologous genes for these three non-colinear genes are present in the *T. halophila* genome, if either gene contributes to the survival of the species, then either or both genes are likely to be present. Among these non-colinear genes, a highly degraded gene fragment of At_a (AT2G01920, Oxygen-evolving protein-coding gene) was identified in the orthologous position of the *T. halophila* sequence. Therefore, it appears that gene disruption by an unknown mechanism has also influenced the evolution of this region. With the sequencing of the *T. halophila* genome, it will be possible to determine if a redundant copy of At_a exists elsewhere in the *T. halophila* genome.

LTR retrotransposons are major contributors to local size variation in the SOS1 orthologous region in T. halophila

Retrotransposable elements have been shown to affect genome structure and evolution in a number of ways leading to genome size variation, genome rearrangement and changes in the regulation of gene expression [25]. Five LTR retrotransposons were identified in the *T. halophila* genome while none were found in this region in *A. thaliana*. All of these elements are located in intergenic regions and account for 15.2% of the total BAC sequence (Table 1). The genome size of *T. halophila* is approximately 240 Mb, which is approximately twice that of *A. thaliana* (125 Mb). LTR retrotransposons have been involved in plant genome size variation due to their copy-and-paste mode of transposition via RNA intermediates [25]. In the grass family, LTR retrotransposons are estimated to make up 14% of the rice genome (430 Mb) [26], 50 to 60% of the maize genome (2500 Mb) [27] and over 70% of the barley genome (4800 Mb) [28]. In the *Brassicaceae*, Class I elements, which include both LTR and non-LTR retrotransposons, comprise 14% of the *B. oleracea* genome (600 Mb) and 4% of the *A. thaliana* genome (125 Mb) [29]. Though the average gene size in *T. halophila* is slightly larger than in *A. thaliana*, this difference did not appear to make a significant contribution to the overall genomic increase. Rather, the average gene density is significantly lower in *T. halophila* indicating that the intergenic regions are responsible for the genomic increase. Therefore, these five LTR retrotransposons in *T. halophila* are major contributors to the size increase found in the *SOS1* region. The insertions of five non-orthologous LTR retrotransposons in the *T. halophila* genome are estimated to have taken place less than 1.5 MYA. Based on estimates from this analysis suggesting that *A. thaliana* and *T. halophila* shared a common ancestor approximately 11.8 MYA (Table 2) prior to speciation, the expansion of the *T. halophila* genomic sequence around the *SOS1* locus appears to have taken place via insertion of LTR retrotransposons after its divergence from *A. thaliana*. Based on the observation that approximately 15.2% of the *T. halophila* sequence around the *SOS1* locus is composed of LTR transposable elements, estimates of the contribution of these elements to the whole *T. halophila* genome would be ~36.5 Mb (240 Mb × 0.152).

ThSOS1 shares similar gene structure and functional domains with AtSOS1 but exhibits a different pattern of SSRs in the 5′ upstream region

ThSOS1 has a similar gene structure to *AtSOS1* and contains 11 transmembrane domains and a cyclic nucleotide-binding site, indicating that *ThSOS1* likely performs a function similar to *AtSOS1*. One feature that distinguishes the two *SOS1* genes is the presence of several SSRs in the 5′ upstream region of *ThSOS1*. Among three SSRs found within 540 bp of 5′ upstream sequence (Fig. 3), the $(CTT/GAA)_n$ repeat contains sequence similar to the TCA-element (TCATCTTCTT) which has been identified as a salicylic acid-responsive element in plants [30]. Zhang et al. [31] reported that 70 to 80% of CTT/GAA-associated genes in *A. thaliana* are regulated by salicylic acid. $(CTT)_{18}$ was identified in the 5′ upstream region 49 bp from the *ThSOS1* translational start site. *AtSOS1* and *OsSOS1* also have CTT repeats in the 5′ UTR, but with very short units of $(CTT)_3$ and $(CTT)_4$, respectively. Zhang et al. [31] reported that RT-PCR analyses resulted in different expression patterns of $(CTT)_n$ or $(GAA)_n$-associated genes in *A. thaliana* in response to salicylic acid treatment. $(CTT)_4$ showed a down-regulated transcription pattern while repeats greater than $(CTT)_5$ showed constant levels of mRNA up to 48 h after treatment. This finding indicated that the number of repeat units affected the transcription or stability of $(CTT)_n/(GAA)_n$-associated genes. The large number of CTT repeats might be a *ThSOS1*-specific feature affecting the synthesis or stability of *ThSOS1* transcripts. So far, genomic sequences of *SOS1* from *Arabidopsis lyrata* and *Oryza sativa* show no apparent CTT repeats longer than four copies (Nah et al. unpublished). Whether $(CTT/GAA)_n$ repeats in the *SOS1* locus respond to salicylic acid and whether salicylic acid is involved in salt tolerance remain to be determined. To date, SSRs in plants have been mainly used for marker applications, although a few examples of functional SSRs have been reported [32,33]. This study has identified an SSR as a possible *cis*-acting element with the potential to differentially regulate *SOS1* in *T. halophila* and *A. thaliana*. The $(TA)_n$ and $(TCA)_n$ repeats associated with *ThSOS1* could also be potential *cis*-acting elements.

Materials and methods

Construction of T. halophila bacterial artificial chromosome library

All BAC library construction protocols were as previously described [34]. Briefly, megabase-size DNA was isolated from *T. halophila* leaf tissue (accession Shandong) using standard procedures. The DNA was partially digested with HindIII, sized-selected on a CHEF gel, the gel purified and then ligated with HindIII digested pBeloBAC11, followed by *E. coli* transformation. Recombinant clones were robotically picked and arrayed into 384-well plates and archived at −80 °C. The BAC library, hybridization filters, and individual clones are available upon request from the Arizona Genomics Institute (www.genome.arizona.edu).

BAC clone selection and sequencing

To obtain a BAC clone containing the *ThSOS1* locus, two regions of a putative *ThSOS1* gene were used as probes. The first probe (1513 bp) was generated by PCR amplification using *T. halophila* genomic DNA as a template and primers designed to the *A. thaliana SOS1* gene (*AtSOS1*) and included 201 bp of the first exon and 1312 bp of upstream sequence. The second probe (1405 bp) was a gene specific probe derived from the 3′ end of a *ThSOS1* cDNA and was designed to distinguish *ThSOS1* from the rest of the Na^+/H^+ antiporter gene family members in *A. thaliana*. These probes were radio-labeled with ^{32}P and used to probe colony hybridization filters derived from a HindIII BAC library from *T. halophila* (accession Shandong) constructed by the Arizona Genomics Institute. Eleven BAC clones were identified from this screen and confirmed by colony PCR with *ThSOS1*-specific primers. The eleven BAC clones were end sequenced using standard protocols [35] and the derived sequences were used to select a BAC clone (FJ386403) that was predicted to contain the *ThSOS1* gene approximately in the middle of the BAC, based on BLAST searches against the *A. thaliana* genome. DNA from the FJ386403 BAC clone was randomly sheared using a Hydroshear (GeneMachines), end-repaired and size-selected 2 to 5 kb inserts were ligated into the pBlueScript KS+ vector (Stratagene) to construct a shotgun library as previously described [36]. Shotgun clones were bi-directionally sequenced using T7 (5′-TAATACGACTCACTATAGGG-3′) and T3 (5′-AATTAACCCTCACTAAAGGG-3′) primers with ABI (Applied Biosystems) Big Dye Terminator 3.1 Chemistry on ABI 3730 XL automated DNA sequencers. Base identification and quality assessments were made using PHRED [37,38]. Shotgun reads (2252 reads) were assembled with PHRAP (http://www.phrap.org/) and edited with CONSED [39] for graphic display of assembly and for sequence finishing. Sequence gaps were filled using a combination of bacterial transposon-mediated [40] and primer walking methods as described previously [35]. The final sequence assembly had an error rate of less than 1 in 10,000 bp. The sequence has been deposited to the NCBI GenBank with accession number of FJ386403.

BAC sequence analysis and annotation

Software programs Dotter [41], PipMaker [42] and ACT [43] were used for the FJ386403 BAC-to-*A. thaliana* genome sequence alignment and analysis. BLASTN, BLASTP, TBLASTN, BLASTX and BLAST2 analyses were also used as required (http://www.ncbi.nlm.nih.gov/BLAST/).

For gene annotation of FJ386403 (193,021 bp), a gene prediction program and EST and protein databases were used. For gene prediction, FGENESH with the *A. thaliana* training set (http://sun1.softberry.com) was chosen. Refinement of gene structure used the *A. thaliana* EST databases from The Institute for Genomic Research (TIGR) (http://compbio.dfci.harvard.edu/tgi/plant.html) and from The Arabidopsis Information Resource (TAIR) (ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/), in addition to the *Arabidopsis* full-length cDNA collection [44]. For functional annotation, protein databases obtained from SwissProt (http://www.ebi.ac.uk/swissprot/), The National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/Genbank/) and TAIR were used. Repeat elements were identified using *Arabidopsis* and other plant repeat databases from the Genetic Information Research Institute (GIRI, http://www.girinst.org/repbase/update/index.html). The automated annotation results were retrieved in an XML file for display and manual annotation in the Apollo Genome Annotation Curation Tool (Version 1.6.5) [45]. When predicted genes matched additional EST databases from species other than *A. thaliana* with an *E*-value less than E^{-20}, they were considered actual genes. Hypothetical genes and transposons were not considered as genes.

Both DNA transposons and retrotransposons were identified by a combination of repeat database searches and structural confirmation. LTR retrotransposable elements were identified by finding pairs of duplicated regions as tentative LTRs using CROSS_MATCH (www.phrap.org). Identification of internal coding sequences between two LTRs and 4 to 6 bp of tandem direct repeats at the 5′ TG and 3′ CA regions of LTRs were used as signatures of LTR structures. Simple Sequence Repeats (SSRs) were searched for separately using the SPUTNIK server (http://cbi.labri.fr/outils/Pise/sputnik.html).

Evolutionary analysis

Investigation of functional constraint between orthologous genes was performed using a Ka/Ks test. A Perl script was used that takes a dataset of cDNA sequences, checks for the absence of stop codons, aligns them in translated sequences, returns the alignments into cDNA sequences to estimate nonsynonymous (Ka) and synonymous (Ks) substitutions using Maximum likelihood (ML) methods of the PAML package [46]. Divergence time between *T. halophila* and *A. thaliana* was estimated using the Ks values of 17 orthologous gene pairs that were calculated based on the PAML package. A mutation rate of 1.5×10^{-8} mutations/site/year was used as described by Koch et al. [5]. Phylogenetic trees were built using amino acid sequences from coding regions. The neighbor joining method under the Poisson correction model was used in MEGA3 with Bootstrap values of 1000. Insertion times of LTR retrotransposable elements were estimated based on the distance between pairs of LTRs. DNA sequence from both LTRs from individual LTR retrotransposable elements were aligned using ClustalW and the distance between them was estimated using the Kimura-2-parameter model implemented in MEGA3 [47]. For estimation of LTR insertion time, the average substitution rate of 1.3×10^{-8} mutations/site/year was used [48].

Acknowledgments

We thank members of the Arizona Genomics Institute's Resource Center, Sequencing, Finishing and Bioinformatics groups for generating and processing the high-quality sequence data used for this analysis, and members of the Wing and Schumaker laboratories for helpful comments on the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2009.05.007.

References

[1] B. Jacoby, Mechanisms involved in salt tolerance of plants, in: M. Pessarakli (Ed.), Handbook of Plant and Crop Stress, E Marcel Dekker Inc., New York, 1999, pp. 97–123.
[2] J.-K. Zhu, Plant salt tolerance, Trends Plant Sci. 6 (2001) 66–71.
[3] G. Inan, et al., 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 (2004) 1718–1737.
[4] Y.W. Yang, K.N. Lai, P.Y. Tai, W.H. Li, Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages, J. Mol. Evol. 48 (1999) 597–604.
[5] M.A. Koch, B. Haubold, T. Mitchell-Olds, Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (*Brassicaceae*), Mol. Biol. Evol. 17 (2000) 1483–1498.
[6] S.I. Warwick, I.A. Al-Shehbaz, R.A. Price, C. Sauder, Phylogeny of *Sisymbrium* based on ITS sequences of nuclear ribosomal DNA, Can. J. Bot. 80 (2002) 1002–1017.
[7] I.A. Al-Shehbaz, M.A. Beilstein, E.A. Kellogg, Systematics and phylogeny of the *Brassicaceae* (*Cruciferae*): an overview, Pl. Syst. Evol. 259 (2006) 89–120.
[8] M.E. Schranz, B.H. Song, A.J. Windsor, T. Mitchell-Olds, Comparative genomics in the *Brassicaceae*: a family-wide perspective, Curr. Opin. Plant Biol. 10 (2007) 168–175.
[9] J.-K. Zhu, Regulation of ion homeostasis under salt stress, Curr. Opin. Plant Biol. 6 (2003) 441–445.
[10] V. Chinnusamy, K. Schumaker, J.-K. Zhu, Molecular genetic perspectives on cross-talk and specificity in abiotic stress signaling in plants, J. Exp. Bot. 55 (2004) 225–236.
[11] H. Shi, M. Ishitani, C. Kim, J.-K. Zhu, The *Arabidopsis thaliana* salt tolerance gene *SOS1* encodes a putative Na^+/H^+ antiporter, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 6896–6901.
[12] Q.S. Qiu, Y. Guo, M.A. Dietrich, K.S. Schumaker, J.-K. Zhu, Regulation of SOS1, a plasma membrane Na^+/H^+ exchanger in *Arabidopsis thaliana*, by SOS2 and SOS3, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 8436–8441.
[13] R. Vera-Estrella, B.J. Barkla, L. Garcia-Ramirez, O. Pantoja, Salt stress in *Thellungiella halophila* activates Na^+ transport mechanisms required for salinity tolerance, Plant Physiol. 139 (2005) 1507–1517.
[14] D.-H. Oh, et al., Sodium stress in the halophyte *Thellungiella halophila* and transcriptional changes in a *thsos1*-RNA interference line, J. Int. Plant Biol. 49 (2007) 1484–1496.
[15] T. Taji, et al., Comparative genomics in salt tolerance between *Arabidopsis* and *Arabidopsis*-related halophyte salt cress using *Arabidopsis* microarray, Plant Physiol. 135 (2004) 1697–1709.
[16] S. Kant, P. Kant, E. Raveh, S. Barak, Evidence that differential gene expression between the halophyte, *Thellungiella halophila*, and *Arabidopsis thaliana* is responsible for higher levels of the compatible osmolyte proline and tight control of Na^+ uptake in *T. halophila*, Plant Cell Environ. 29 (2006) 1220–1234.
[17] G.G. Loots, et al., Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons, Science 288 (2000) 136–140.
[18] M. Mohrs, et al., Deletion of a coordinate regulator of type 2 cytokine expression in mice, Nat. Immunol. 2 (2001) 842–847.
[19] S. Huang, et al., Comparative genomics enabled the isolation of the *R3a* late blight resistance gene in potato, Plant J. 42 (2005) 251–261.
[20] M. Gao, G. Li, B. Yang, W.R. McCombie, C.F. Quiros, Comparative analysis of a

Brassica BAC clone containing several major aliphatic glucosinolate genes with its corresponding *Arabidopsis* sequence, Genome 47 (2004) 666–679.

[21] *Arabidopsis* Genome Initiative, Analysis of the genome sequence of the flowering plant *A. thaliana*, Nature 408 (2000) 796–815.

[22] D. Swarbreck, et al., The Arabidopsis Information Resource (TAIR): gene structure and function annotation, Nucleic Acids Res. 36 (2008) D1009–D1014 Database issue.

[23] K.H. Wolfe, W.H. Li, P.M. Sharp, Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs, Proc. Natl. Acad. Sci. U. S. A. 84 (1987) 9054–9058.

[24] J.D. Palmer, L.A. Herbon, Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence, J. Mol. Evol. 28 (1988) 87–97.

[25] A. Kumar, J.L. Bennetzen, Plant retrotransposons, Annu. Rev. Genet. 33 (1999) 479–532.

[26] N. Jiang, Z. Bao, X. Zhang, S.R. Eddy, S.R. Wessler, Pack-MULE transposable elements mediate gene evolution in plants, Nature 431 (2004) 569–573.

[27] B.C. Meyers, S.B. Tingey, M. Morgante, Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome, Genome Res. 11 (2001) 1660–1676.

[28] C.M. Vicient, et al., Retrotransposon *BARE-1* and its role in genome evolution in the genus Hordeum, Plant Cell 11 (1999) 1769–1784.

[29] X. Zhang, S.R. Wessler, Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 5589–5594.

[30] A.P. Goldsbrough, H. Albrecht, R. Stratford, Salicylic acid-inducible binding of a tobacco nuclear protein to a 10 bp sequence which is highly conserved amongst stress-inducible genes, Plant J. 3 (1993) 563–571.

[31] L. Zhang, et al., Conservation of non-coding microsatellites in plants: implication for gene regulation, BMC Genomics 7 (2006) 323.

[32] S. Bao, H. Corke, M. Sun, Microsatellites in starch-synthesizing genes in relation to starch physicochemical properties in waxy rice (*Oryza sativa* L.), Theor. Appl. Genet. 105 (2002) 898–905.

[33] R.J. Hulzink, et al., The 5′-untranslated region of the *ntp303* gene strongly enhances translation during pollen tube growth, but not during pollen maturation, Plant Physiol. 129 (2002) 342–353.

[34] M. Luo, R.A. Wing, An improved method for Plant BAC library construction, in: E. Grotwold (Ed.), Methods in Molecular Biology: Ed 1 Vol 236. Plant Functional Genomics: Methods and Protocols, Humana Press, Totowa, 2003, pp. 3–19.

[35] M. Luo, et al., Utilization of a zebra finch BAC library to determine the structure of an avian androgen receptor genomic region, Genomics 87 (2006) 181–190.

[36] C.E. Grover, H. Kim, R.A. Wing, A.H. Paterson, J.F. Wendel, Incongruent patterns of local and global genome size evolution in cotton, Genome Res. 14 (2004) 1474–1482.

[37] B. Ewing, P. Green, Base-calling of automated sequencer traces using Phred. II. Error probabilities, Genome Res. 8 (1998) 186–194.

[38] B. Ewing, L. Hillier, M.C. Wendl, P. Green, Base-calling of automated sequencer traces using Phred. I. Accuracy assessment, Genome Res. 8 (1998) 175–185.

[39] D. Gordon, C. Abajian, P. Green, Consed: a graphical tool for sequence finishing, Genome Res. 8 (1998) 195–202.

[40] S. Haapa, et al., An efficient DNA sequencing strategy based on the bacteriophage *mu* in vitro DNA transposition reaction, Genome Res. 9 (1999) 308–315.

[41] E.L. Sonnhammer, R. Durbin, A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis, Gene 167 (1995) GC1–10.

[42] S. Schwartz, et al., PipMaker-a web server for aligning two genomic DNA sequences, Genome Res. 10 (2000) 577–586.

[43] T.J. Carver, et al., ACT: the Artemis Comparison Tool, Bioinformatics 21 (2005) 3422–3423.

[44] M. Seki, et al., Functional annotation of a full-length *Arabidopsis* cDNA collection, Science 296 (2002) 141–145.

[45] S.E. Lewis, et al., Apollo: a sequence annotation editor, Genome Biol. 3 (2002) RESEARCH 0082.1–0082.14.

[46] Z. Yang, PAML: a program package for phylogenetic analysis by maximum likelihood, Comput. Appl. Biosci. 13 (1997) 555–556.

[47] S. Kumar, K. Tamura, M. Nei, MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment, Brief Bioinform. 5 (2004) 150–163.

[48] J. Ma, K.M. Devos, J.L. Bennetzen, Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice, Genome Res. 14 (2004) 860–869.