A whole-genome shotgun assembly for genome characterization of the common ice plant (Mesembryanthemum crystallinum L.)

Ryoma Sato
Kyushu University
Yuri Kondo
Kyushu University
Sakae Agarie (✉ agarie@agr.kyushu-u.ac.jp)
Kyushu University

Article

Keywords:

Posted Date: September 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2013540/v1

Abstract

The common ice plant (*Mesembryanthemum crystallinum* L.) is an annual herb belonging to the genus Mesembryanthemum family of the family Aizoaceae. Here, we performed shotgun genome paired-end sequencing using the Illumina platform to determine the genome sequence of the ice plants. A draft genome was generated with a total length of 286 Mb corresponding to 79.2% of the estimated genome size (361 Mb), consisting of 49,782 contigs. It encompassed 93.49% of the genes of terrestrial higher plants, 99.5% of the ice plant transcriptome, and 100% of known DNA sequences. In addition, 110.9 Mb (38.8%) of repetitive sequences and untranslated regions, 971 tRNA, and 100 miRNA loci were identified, and their effects on stress tolerance and photosynthesis were investigated. Overall, 35,702 protein-coding regions were identified in the genome, of which 56.05–82.59% were annotated and used in domain searches and gene ontology (GO) analyses. The functional characterization using ice plant draft genome is a fundamental result. It can be helpful to elucidate the mechanism of growth promotion and reversible conversion of the photosynthetic type from C3 to CAM in the presence of NaCl. Further, these data could be used in the creation of novel, extremely salt-tolerant crops.

1. Introduction

Soil salinity is one of the most detrimental types of abiotic stress. Osmotic and ionic stresses cause significant decreases in plant growth. Developing a wide range of strategies for adapting to and mitigating NaCl stress is required to address the negative impacts of salinity. Efficient resource management and crop improvement will help overcome the salinity-induced damages to agricultural production. *Mesembryanthemum crystallinum* L. or common ice plant is an annual plant of the family Aizoaceae, native to South Africa. This plant survives in the presence of a high salt concentration, even higher than that of seawater (about 500 mM) (salt-tolerance), and can grow under moderate salinity, up to 200 mM NaCl, wherein the growth and development of most crops are severely inhibited (halophilism)[1]. Also, it can convert its photosynthetic mode from C_3 to CAM under severe salt stress and drought stress[2]. For the past half-century, the common ice plant has been frequently used as a model for elucidating the mechanisms of salt stress tolerance and photosynthetic conversion in response to salt and drought stresses. Recently, the molecular mechanisms underlying these phenomena have been elucidated using epigenetics, transcription, post-transcription, translation, post-translation analysis focusing on specific proteins, mitochondria, and chloroplasts. Large-scale gene expression profiling has been conducted using expression sequence tags (ESTs), microarrays, and next generation sequencing (NGS).

Although the functions of the large number of differentially regulated genes in the presence of NaCl are known, the entire genome has not been available, delaying the elucidation of whole-genome functions. The genome sequence contains protein-coding genes, non-transcribed regions, including promoters and terminators, together with untranslated regions, including micro-RNAs (miRNAs). The untranslated regions influence gene expression by affecting the mRNA stability and translational efficiency[3].

Information regarding the genome sequences' biological functions facilitates a comprehensively understanding the epigenetic and transcriptional regulatory mechanisms of gene expression. Sequencing the common ice plant's genome is essential to clarify the whole-genome's response to NaCl, which can help genome editing and breeding studies to generate superior varieties.

Short read de novo genome assembly has been mainly used for organisms with small genomes and few repetitive sequences, such as microorganisms. New technologies have increased read lengths and paired-end and mate-pair sequencing have been established, enabling short-read NGS to also be used for the de novo assembly of large eukaryotic genomes. Short-read sequencing costs less than the long-read sequencing obtained using third and fourth NGS. Several software programs for *de novo* genome assembly for short reads have been developed. The algorithm for genome assembly (ALGA) is the newest assembler, based on an overlapping graphs model, which can generate more accurate results than conventional software using the de Bruijn graphs model[4].

In this study, we constructed the ice plant genome using easy-to-start applications such as ALGA to accelerate genome analysis. We investigated the characteristics of the genome, clarifying the repetitive sequences, tRNAs, and miRNAs (genomic regions and precursors), and identified gene regions using various software and web tools. This is the first report of whole-genome analysis of the common ice plant. Our results indicate the involvement of translated and untranslated regions in the regulatory processes of salt tolerance and photosynthetic conversion under stress in the ice plant.

2. Results

2.1 Genome sequencing and *de novo* genome assembly

Short insert reads data (300 bp) (Fig. S1-(A)) with a ratio of unique to duplicate reads of 31:19 was obtained by removing a total of 30 Gb (∼30×) of erroneous reads of raw paired-end data from the Illumina platform (BioProject: PRJDB13817; BioSample: SAMD00508673) (Table S1). The *M. crystallinum* genome size was estimated to be 358 to 364 Mb, with very low heterozygosity (about 0.090%) following an analysis of the frequency of 21 and 25-mers, using GenomeScope2.0 (Fig. S1-(B) and (C)). The *M. crystallinum* final draft assembly included 286 Mb in 49,782 scaffolds with a scaffold N50 of 10,562 bp (Table S2). The BUSCO tool revealed 1509 (93.49%) of 1,614 embryophyte library core genes, with 1,223 (75.77%) of these being 'Complete' matches in the genome. Although the completeness and contiguity of the genome were lower than those of genome assemblies of other plants with long scaffolds, they were greater than the shotgun assembled *M. australis* and *S. chaucha* genomes (Fig. S2). Around 24,081 (99.5%) of the 24,204 transcripts from the transcriptome assembly of *M. crystallinum* leaves[5], and all 135 DNA sequences registered in the NCBI, were aligned to the assembled genome (Supplementary Dataset S1).

2.2 Search and classification of repetitive regions

In the 286.0 Mb *M. crystallinum* genome, 2,423 distinct repetitive sequences families, accounting for 110.9 Mb (38.8%) of the genome, were identified using custom repeat libraries and Repbase[6]. In decreasing order of frequency, the annotated repetitive elements, were unclassified 78.0 Mb (27.27%), retroelements 21.9 Mb (7.64%), long interspersed nuclear elements (LINE) 12.5 Mb (4.37%), long terminal repeats (LTR) 9.35 Mb (3.27%), and simple repeats 7.26 Mb (2.54%). some retroelements were classified into subfamilies, including L1/CIN4 12.4 Mb (4.34%) and RTE/Bov-B 0.85 Mb (0.03%) in the LINE, and Ty1/Copia 4.90 Mb (1.71%) and Gypsy/DIRS1 4.35 Mb (1.52%) in the LTR (Table 1).

Table 1
Classification results of repetitive sequences in the ice plant genome.

Group	Number of elements	Length occupied, [bp]	Percentage of sequence, [%]
Retroelements [1]	42,672	21,857,207	7.64
LINEs [2]:	25,415	12,509,921	4.37
RTE/Bov-B	278	84,896	0.03
L1/CIN4	25,137	12,425,025	4.34
LTR elements [3]:	17,257	9,347,286	3.27
Ty1/Copia	9,514	4,904,145	1.71
Gypsy/DIRS1	7,528	4,345,997	1.52
DNA transposons [4]	5,725	2,789,199	0.98
hobo-Activator	728	285,258	0.10
Tc1-IS630-Pogo	237	160,874	0.06
Tourist/Harbinger	590	263,259	0.09
Rolling circles	121	112,702	0.04
Unclassified:	392,582	77,986,137	27.27
Total interspersed repeats:		102,632,543	35.88
Simple repeats:	137,610	7,255,343	2.54
Low complexity:	19,059	911,926	0.32

[1] Retroelements: DNA sequences derived from viruses.

[2] LINEs: Long interspersed nuclear elements.

[3] LTR elements: Retrotransposons with long terminal repeat.

[4] DNA transposons: DNA sequences moving through the genome.

2.3 Detection of tRNA and miRNA coding genes from the genome

A total of 971 tRNAs, excluding pseudogenes, were detected in the assembled genome, and were sorted into several groups based on codon designation. The codon with the most abundant tRNA was isoleucine and the least was tryptophan (Fig. S3). The number of tRNAs was as follows: *Arabidopsis* 585, rice 505, poplar 505, tomato 723, horseradish 500, potato 736, grape 391, and soybean 700. Interspecific comparisons using the Smirnov-Grubbs outlier test and focusing on these eight species indicated that the abundance of isoleucine was significantly highest and that of tryptophan was significantly lowest ($P < 0.05$) (Fig. 1 and Table S3).

In addition, miRNAs loci were identified from the genome with reference to the Rfam database, to obtain miRNA profiling independent of their expression levels. MiRNAs are 21 to 24 nt molecules that regulate post-transcriptional mRNA modification, playing important roles in plant growth and tolerance to environmental stress. One hundred miRNA loci were identified and categorized into 25 families. The RNA family with the largest number of loci was MIR169 (25), followed by mir-399 (16), MIR159 (8), and mir-166 (7). mRNAs targeted by miRNA families were predicted (Table S4). For instance, MIR169 family miRNAs were presumed to bind to mRNAs encoding nuclear factor gamma subunit A (NF-YA)[7]. Overall, thirteen types of 25 miRNA families were likely to target mRNAs encoding transcription factors: MYB33, MYB65, HD-ZIP, WRKY, AP2-like, NAC, ARFs, IAR3, ARF16, OsSPL14, SPL, GRF2, and HLH. The rest of the targeted mRNAs are anticipated to have functions in processes such as miRNA maturation, mRNA cleavage, or metal binding.

2.4 Gene prediction and annotation

Genes (34,223), coding sequences (35,702), and amino acid regions (35,702) were predicted from the soft-masked draft *M. crystallinum* scaffolds *ab initio* using a homology-based pipeline in BRAKER2 using transcriptome data (Table S4). The representative value on bases showed that coding sequence regions cover at least 10.6% (30.4 Mb) of the total genome sequence. In comparison to several databases on 25 plant species genes registered in PGDBj[8] (Last accessed in March 2022), the ice plants' genes were as abundant as those of *Sorghum bicolor* and *Arabidopsis lyrate*. Additionally, summarized data indicated that the *M. crystallinum* gene number was 16 times larger than those of *S. bicolor* and *A. lyrate*, equivalent to about 27.6% of the number of genes of *Triticum aestivum* (bread wheat) and 3.31-fold greater than that of *Pyropia yezoensis* (bangia) (Fig. S4). Each translated protein sequence was used in a BLASTP search with the DIAMOND program[9] against four kinds of protein sequence databases. In order of the proportion of homologous amino acid sequences identified, they were NCBI-non-redundant (82.59%), poplar (70.65%), TAIR10 (65.39%), and swissprot (56.05%) (Table 2) (Supplementary Dataset S2). To simplify gene ID conversion to GO terms, the results, including TAIR ID, were used in the functional estimation.

Table 2
Summary of predicted genes in the common ice plants (M. crystallinum) scaffolded sequence.

	Databases			
	NCBI	**Swiss-prot**	**TAIR10**	***Populus trichocarpa***
Annotated genes	29,485	20,012	23,346	25,223
Coverage[(1)], [%]	82.59	56.05	65.39	70.65

[(1)] Coverage refers to the ratio of homologous genes to 35,702 different amino acid sequences.

2.5 Functional estimation and comparison of genomes

A Pfam domain search based on the Pfam[10] database identified 3,703 domains in 23,521 (97.1%) genes. The most frequently occurring domain was the protein kinase domain (PKinase), at 2.18%, followed by a domain of unknown function (DUF) 4238 (1.85%), reverse transcriptase (RVT)_1 (1.45%), PPR domain-containing protein (PPR)_2 (1.42%), and protein tyrosine and serine/threonine kinase (PK_Tyr_Ser-Thr) (1.18%) (Supplementary Dataset S3). The PKinase family was further classified into 94 kinase families using iTAK[11]. The top 30 kinase families with the largest number of ice plant genes are shown in descending order in Fig. 2. Compared to the other four plant species, the proportion of 12 families was significantly higher ($P < 0.05$), and eight families—DUF4238, RVT_1, RVT_2, RVT_3, Retrotrans_gag_2, zf_RVT, Retrotrans_gag_3, Retrotrans_gag—contained retroelement domains that could be attributed to a retrotransposable element (Fig. 3). The annotated genes were assigned to GO classifications based on TAIR ID in three groups —biological process (BP), cellular component (CC), and molecular function (MF)—and were categorized into 403 GO terms using the gene functional classification tool in the DAVID web service. The proportion of genes assigned to 94 GO terms did not differ significantly among five plant species ($P > 0.05$) (Fig. S5 to S7), indicating that they are essential to plant survival. These findings confirmed that the ice plant genome constructed in this study contained some universally present genes. Eighteen GO terms were identified only from ice plants, although the number of genes was small (Table S6), involving virus resistance, pollen tube development, and fat biosynthesis (BP); cytoplasmic vesicle and “soluble NSF attachment protein receptor” (SNARE) (CC); and *O*-acyltransferase for transferring fatty acids (MF).

3. Discussion

M. crystallinum is a model plant for investigating halophilism, salt tolerance, and CAM photosynthesis, given its multiple stress tolerance and stress-induced photosynthetic conversion. In this research, the common ice plant’s genome sequence was assembled from short-read sequences and used for the entire elucidation of the genome’s contents in detail for the first time. This genomic resource covers all protein-coding, non-transcribed, and untranslated regions. Genomic functional analysis data about *M. crystallinum* provides new insights into the molecular mechanisms underlying the plant’s adaptation to NaCl stress and its ability to convert the photosynthetic mechanism.

We found the repetitive *M. crystallinum* sequences to occupy 110.9 Mb (38.8%) of the genome. Advances in genomics over several decades have revealed that most repetitive sequences play essential roles in regulating the gene expression related to stress and photosynthetic responses in higher plants. According to recent studies involving *Arabidopsis*, tomato, and mangrove species, transposable elements, comprising many repetitive sequences, were highly expressed under heat, salt, and intense light stresses. These results suggested that it affected the expression levels of nearby genes for transcriptional factors, including *DREB*, *NAC*, *MYB*, *AP2/ERF*, *NF-Y*, and *Abscisic acid 8'-hydroxylase*[12,13]. Further studies indicated that *cis*-regulatory motifs associated with C4 photosynthesis up to at least 669 and the non-coding RNAs regulating methyltransferases expression levels are derived from transposable elements[14,15]. Transposable element expression is suppressed by cytosine methylation in DNA sequences, chromatin remodeling, and degradation by small interfering RNA (siRNA)[16]. The terrestrial common ice plant is suggested to have repetitive sequences with similar effects on gene expression regulation.

Two kinds of representative small non-coding RNAs were found in the ice plant genome—971 tRNAs and 100 miRNA loci—which are anticipated to be relevant to stress reduction and post-transcriptional modification. Generally, a tRNA recruits an amino acid corresponding to its codon, which means that the abundance of a specific tRNA is proportional to that of the relevant amino acid. Some studies have shown the effectiveness of amino acids in metabolism for environmental stress reduction. For example, 5-aminolevulinic acid, a key precursor in porphyrins biosynthesis, including chlorophyll and heme, can alleviate abiotic stresses, including salinity, drought, heat, cold, and UV-B[17]. The Smirnov-Grabs outlier test revealed that the isoleucine-specific tRNA was present at a significantly higher percentage in our data than in eight other plant species investigated in this study. It is the precursor of JA-Ile, the active molecule of the plant hormone jasmonic acid, which has been implicated in pathogen resistance in plants[18]. The least abundant tRNA tryptophan specific, which serves as the melatonin precursor, a signaling molecule that regulates responses to abiotic stress, such as water shortage[19]. These results suggest that the abundance of amino acids in the ice plant may differ from those in the other eight plants, indicating that it has a different stress tolerance mechanism.

Some miRNAs identified in the ice plant's genome appeared to be key small molecules in the stability of mRNAs coding for epigenetic and transcription-related factors. NF-YA were targeted by 31 MIR169 loci known to integrally regulate gene expression by maintaining histone acetylation in soybean[20], or binding to circadian rhythm-related elements, including the "CCAAT" motif in *Arabidopsis*[21,22]. Several miRNA-targeting transcription factors were associated with salt tolerance (*HLH*, *SPL*, *HD-ZIP*)[23–25] or CAM-type photosynthesis (*WRKY*, *AP2*, *MYB*, *NAC*)[26–28]. All target gene families were found in the protein family collection in the ice plant genome, except for *SPL* and *lectin receptor kinase* (see Supplementary Dataset S1), indicating that an antagonistic relationship between miRNAs and mRNAs underlies the stress tolerance and photosynthetic conversion mechanisms of the ice plants. Additional miRNA sequence information is expected to provide more accurate data and form the basis for testing these assumptions.

The richest PKinase subfamily was “receptor-like kinase/Pelle, DUF26, SD-1, LRR-VIII and VWA, a moss-specific new RLK subfamily” (RLK-Pelle_DLSV), containing primarily receptor-type kinases, which was consistent with the transcriptome profiling in a halophyte, *Nitraria sibrica*[29]. It has been assumed to be involved in cell wall biosynthesis, adhesion, and developmental regulation. For instance, WAK, the second most frequent PKinase in the ice plant genome, has been reported to control cell wall expansion, metal resistance, and pathogen resistance[30]. The common ice plants show halophilism, or salt tolerance; a detailed study may help to shed light on the mechanism of this tolerance from the perspective of phosphorylation. In contrast to the rare PKinase, the richness of retrotransposon-derived domains (reverse transcriptase and gag genes), involved in RNA packaging and the replication cycle[31], was apparent in ice plants compared to other plant species. A recent study suggested a human retrotransposon-derived imprinted gene, *paternally expressed gene 10* (*PEG10*), mediates cellular proliferation and inhibits apoptosis[32]. However, it remains unclear what effects these types of proteins have on the plants' physiology. Our latest experiment demonstrated that the ice plant’s cell cycle-related genes were upregulated in the presence of 100 mM NaCl[33], possibly implying an impact on cell division by retrotransposon-derived proteins. Lipases, transferases, and phosphatases were abundant, and transcription factors such as Myb, HLH, and AP2 were scarce in the genome of the common ice plant. Some details about the promotion of metabolism and gene expression by these enzymes and transcription factors in the presence of NaCl[34,35] are known, but it is not yet clear whether they mutually act. Elucidation of these protein interactions by transcriptome and interactome analysis may provide crucial evidence about their unknown functions.

At last, comparing the gene functions among the genomes of five plant species —ice plant, *Arabidopsis*, rice, maize, and poplar—based on their gene counts, eighteen gene functions were possessed only by the ice plant and appeared to be vital for survival. Previous studies (12 reviews and 11 research articles) with sophisticated experimental backgrounds indicated that all gene functions were possibly associated with halophilic and salt-tolerant (photosynthetic conversion) mechanisms. These gene functions were categorized as related to biological defense, growth, reproduction, transcription, post-transcription, and intermembrane transportation; therefore, focusing on homologous ice plant genes with these functions may provide critical insight into the salt-induced growth and photosynthetic systems.

4. Conclusion

We succeeded in assembling the M. crystallinum genome using Illumina PE reads, characterizing the genome, and identifying the potential gene, non-transcriptional and translational regions, and repetitive sequences. This achievement can be regarded as a model case of a genome study using NGS short reads, given its level of success. Analysis revealed that salt tolerance increases with growth, and C3-CAM photosynthetic conversion in the presence of NaCl is probably controlled by both protein-coding genes and potential genomic factors, including transposable elements, tRNAs, miRNAs, and protein kinases. These findings provide new insights into the mechanisms of plant growth under environmental stresses and can be used to develop highly salt-tolerant crops. We hope this study will be a key to the genomic science of the common ice plant.

5. Materials And Methods

All the processes involved in this study are shown in Supplementary Fig S8.

5.1 Plant materials and growth conditions

Seeds of the common ice plant (*Mesembryanthemum crystallinum*) were assigned personally from Dr. John C. Cushman from the University of Nevada and stored under coolness and darkness until use. Originally, wild-type seeds were collected from the plants identified by Dr. Klaus Winter, an expert on the common ice plant, on a coastal cliff at the Mediterranean Sea shore close to Caesarea in Israel (around N32° 29' 43.4", E34° 53' 22.8") in 1978[36]. Three voucher specimens of *M. crystallinum* have been deposited in the Herbarium at the Royal Botanic Gardens Kew (55793.000, K000296094, and K000267571). In this study, our biological materials were recognized as the same plants as those specimens. Experiments, including collecting samples for this study, were conducted in compliance with relevant institutional, national, and international guidelines and laws. The seeds were sown on agar plates as described previously[37], and grown in a growth chamber under 12 h of light and 12 h of darkness at 25 ℃. The two-week-old seedlings were transferred to plastic pots filled with growth medium composed of 50% peat moss, 30% cocopeat, and 20% perlite, specified for the ice plants (Japan Agricultural Cooperatives Ito-shima, Fukuoka, Japan) and irrigated with a nutrient solution of 1.5 g/L OAT House No. 1 and 1.0 g/L No. 2 (OAT Agrio Co., Ltd., Tokyo, Japan) in a greenhouse at Kyushu University for five weeks. The plants were then treated with 0.3% NaCl (w/w) for two weeks, according to the methods published by Agarie et al. (2009)[38]. Approximately 0.6 g of tissue from each leaf was collected, quickly frozen in liquid nitrogen, and stored at − 80 ℃.

5.2 DNA extraction, library construction, and sequencing

Total genomic DNA was extracted from the leaf tissue and purified using MagExtractor™-Plant Genome Nucleic Acid Purification Kits (Toyobo Co., Ltd., Shiga, Japan), according to the manufacturer’s instructions. The DNA samples were fragmented by sonication and used to construct short insert paired-end libraries construction using NEBNext® Ultra™DNA Library Prep Kits for Illumina (New England Biolabs Ltd., Ipswich, MA, USA). Briefly, in the end-repair step, fragmented DNA was phosphorylated at the 5' end and adenylated at the 3' end. During the ligation step, full-length circulated adaptor sequences were ligated to the fragments. After adaptor cleavage, purification and size selection were performed. The indexed PCR products were taken to obtain the final sequencing libraries. The mean insert size for paired-end libraries was 300 bp. The paired-end (2×150 bp) sequencing was conducted on an Illumina NovaSeq 6000 platform (Illumina Inc., San Diego, CA, USA) with a sequencing coverage of 30.

5.3 Clean read preparation and genome size estimation

The mean insert size was calculated using REAPR (v1.0.18)[39], and raw paired-end sequences were filtered based on the frequency of 21-mer sequences using the program Musket (v1.1) [40]. The key parameter values were as follows: musket -omulti output -inorder pair1.fastq pair2.fastq. Sequence reads that appeared rarely or abnormally frequently were removed to obtain clean read data. In the corrected reads, unique and duplicate read numbers in the corrected reads were measured using fastqc (v0.11.9)[41]. The clean data were used for an estimate of genome size as follows. *K*-mers were counted and exported to histogram files using jellyfish (v2.3)[42] [key parameter: jellyfish histo reads.jf]. GenomeScope2.0[43] corresponding key parameters were applied to calculate the genome sizes using *k*-mers lengths of 21 and 25.

5.4 *De novo* genome assembly and quality evaluation

Short reads were assembled using ALGA (v1.0.3)[4] with the default parameter –error-rate = 0.02. long DNA fragments 1 to 10 kb in length were combined, and gaps between them were filled with unknown bases (Ns) using Redundans (v0.14a)[44], a software program for scaffolding, with default parameter values. The completeness of the assembled genome was evaluated based on the content of orthologs in higher plants, using the benchmarking universal single-copy orthologs (BUSCO) program (v5.0)[45]. The lineage dataset was embryophyta_odb10 (creation date: 2020-09-10, number of BUSCOs: 1614). We also searched for core genes in the genome sequences of nine other plant species: *Kewa caespitosa*, *Pharnaceum exiguum*, *Macarthuria australis*, *Solanum chaucha*, *Populus trichocarpa*, *Arabidopsis thaliana*, and *Oryza sativa* using BUSCO. The first three species belong to the same order, Caryophyllales, to which the ice plants belong. Genome information was obtained from the NCBI (see Supplementary Note 1, “Address to genome information”). The number of bases, sequences, sequences in several base number ranges, and maximum base length of the final draft genome sequences were calculated using gVolante (https://gvolante.riken.jp/ Accessed in February 2022; v2.0.0)[46]. BLASTN (v2.2.31+)[47] was used to investigate the number of cDNA sequences identified by transcriptome[5], and registered DNA sequences (Retrieved from https://www.ncbi.nlm.nih.gov/gene/?term=Mesembryanthemum+crystallinum, last accessed February 2022) were aligned to the final assembled genome sequence.

5.5 Detection of repetitive regions

Repetitive sequences were detected, and custom repeat libraries involving transposable elements and long terminal repeat-retro transposons were generated using RepeatModeler2 (v2.0.2)[48] and TEclass (v2.1.3)[49]. Known repeat sequences were detected and classified in the assembled genome sequence with reference to the Repbase library[6] and the custom repeat libraries, using RepeatMasker (v4.1.2-p1) (http://www.repeatmasker.org; last accessed February 2022). The capital letters in genome sequences were replaced with small characters as softmasking.

5.6 Search for genomic sequences coding transfer RNA (tRNA) and micro-RNA (miRNA)

The tRNA genes were identified in the draft common ice plant genome using tRNAscan-SE2.0 (v2.0.9)[50]. The tRNA data of other nine plant species—*Arabidopsis*, rice, tomato, poplar, horseradish, potato, grape, soybean, and coffee tree (robusta species)—were obtained from the PlantRNA database[51]. The percentages of arbitrary tRNAs against the total tRNAs in the genome were calculated and compared to the ice plants' values with the others' ones. Smirnov-Grubbs' outlier tests were performed to select tRNAs more significantly involved. The test statistic T was calculated using the following equation:

$$\mathrm{T} = \frac{(\mathrm{PercentageofarbitrarytRNAsintheiceplant}) - (\mathrm{Samplemeanforallninespecies})}{\sqrt{\mathrm{Samplevariance}}}$$

The miRNA loci in the genome sequence were identified using the cmscan command in infernal (v1.1.4)[52] using Rfam (https://rfam.xfam.org/; last accessed February 2022).

5.7 Gene prediction

The BRAKER2 pipeline (v2.1.5)[53] was used for the prediction of genes in the common ice plant genome. Amino acid sequences were translated from the transcriptome profile reported by Lim et al. (2019)[5] and used as additional reference data for the prediction of genes. BRAKER2 was used with the default parameters (−softmasking). The total sequences, total bases, total amino acids, and N50 were computed based on the resulting fasta-format files containing information about the genes, coding sequences, and amino acids using seqkit (v2.0.0) [54] [key parameter: seqkit stats]. Protein BLAST searches (*E*-value < 1e-5) were conducted using DIAMOND (v2.0.13.151)[9] against the NCBI-non-redundant protein sequences (Retrieved from https://ftp.ncbi.nlm.nih.gov/blast/db/ in March 2022), Uniprot-swissprot (Retrieved from https://www.uniprot.org/ on March 18), Ensemble TAIR10 (Retrieved from http://ftp.ensemblgenomes.org/ on March 2022), and NCBI poplar amino acid sequence databases (Retrieved from https://www.ncbi.nlm.nih.gov/genome/?term=Populus+trichocarpa, last accessed in March 2022).

5.8 Protein domain searches

The protein domains in the genome were identified using the Pfam (v33.1) database[10] with *E*-value < 1e-3, using HMMER (v3.1b2)[55]. The protein databases of rice, maize, and poplar from the NCBI (last accessed February 2022) were used in domain prediction for the purposes of comparison. For a detailed classification of the PKinase family, the iTAK (v18.12) web tool[11] (Last accessed February 2022) was utilized. The ratio of families with a high ratio of genes to total genes in the ice plant was compared with that of the same families in other plants. For statistical analysis, we used Smirnov-Grubbs' outlier tests. The following equation was used to obtain the test statistic T:

$$\mathrm{T} = \frac{(\mathrm{Percentageofarbitraryfamiliesoftheiceplants}) - (\mathrm{Samplemeanforallthreespecies})}{\sqrt{\mathrm{samplevariance}}}$$

Finally, BLASTP was used to compare proteins generated from the ice plant genome and those from *Arabidopsis*, rice, maize, and poplar and renamed TAIR10 ID. These IDs were subjected to gene ontology (GO) enrichment analysis using DAVID (updated in 2022; accessed on March 24)[56] based on a modified Fisher exact probability test with E-value < 0.05.

Declarations

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 21K19120. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Authors and Affiliations

Graduate school of Bioresource and Bioenvironmental Sciences, Kyushu University, 744 Motooka Nishi-ku Fukuoka, 819-0395, Japan

Ryoma Sato, Yuri Kondo

Faculty of Agriculture, Kyushu University, 744 Motooka Nishi-ku Fukuoka, 819-0395, Japan

Sakae Agarie

Contributions

S.A.: Conceptualization, supervision, and funding acquisition. Y.K.: Cultivation, performing experiments, and collecting data. R.S.: Analyzation, investigation, visualization. S.A. and R.S.: Writing the manuscript. All authors took part in revising or critically reviewing the article and gave final approval of the version to be published.

Corresponding Author

Correspondence to Sakae Agarie

Data Availability

The data have been deposited with links to BioProject accession number PRJDB13817 in the DDBJ BioProject database. BioSample metadata are available in the DDBJ BioSample database under accession number SAMD00508673.

Additional Information

Competing interests

The authors declare no competing interests.

References

1. Agarie, S. Possibility of desalinization of saline soils by common ice plant (*Mesembryanthemum crystallinum*). *Trop. Agric. Dev***48**, 294–298 (2004).
2. Adams, P. *et al.* Tansley review No. 97 growth and development of *Mesembryanthemum crystallinum* (Aizoaceae). *Genetics and Breeding***138**, 171–190 (1998).
3. Hughes, T. A. Regulation of gene expression by alternative untranslated regions. *Trends in Genetics***22**, 119–122 (2006).
4. Swat, S. *et al.* Genome-scale *de novo* assembly using ALGA. *Bioinformatics***37**, 1644–1651 (2021).
5. Lim, S. D., Lee, S., Choi, W. G., Yim, W. C. & Cushman, J. C. Laying the foundation for crassulacean acid metabolism (CAM) biodesign: expression of the C4 metabolism cycle genes of CAM in *Arabidopsis*. *Front Plant Sci***10**, 101 (2019).
6. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA***6**, 4–9 (2015).
7. Chiang, C. P. *et al.* Identification of ice plant (*Mesembryanthemum crystallinum* L.) microRNAs using RNA-seq and their putative roles in high salinity responses in seedlings. *Front Plant Sci***7**, 1143 (2016).
8. Asamizu, E. *et al.* Plant genome database Japan (PGDBj): a portal website for the integration of plant genome-related databases. *Plant Cell Physiol***55**, e8 (2014).
9. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods***18**, 366–368 (2021).
10. Mistry, J. *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res***49**, D412–D419 (2021).
11. Zheng, Y. *et al.* iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol Plant***9**, 1667–1670 (2016).
12. Deneweth, J., Van de Peer, Y. & Vermeirssen, V. Nearby transposable elements impact plant stress gene regulatory networks: a meta-analysis in *A. thaliana* and *S. lycopersicum*. *BMC Genomics***23**, 18 (2022).
13. Wang, Y., Dai, A. & Tang, T. Weak effect of Gypsy retrotransposon bursts on *Sonneratia alba* salt stress gene expression. *Front Plant Sci***12**, 830079 (2022).
14. Cao, C., Xu, J., Zheng, G. & Zhu, X. G. Evidence for the role of transposons in the recruitment of cis-regulatory motifs during the evolution of C4 photosynthesis. *BMC Genomics***17**, 201 (2016).
15. Nosaka, M. *et al.* Role of transposon-derived small RNAs in the interplay between genomes and parasitic DNA in rice. *PLoS Genet***8**, e1002953 (2012).
16. Ito, H. Small RNAs and regulation of transposons in plants. *Genes Genet Syst***88**, 3–7 (2013).
17. Tan, S., Cao, J., Xia, X. & Li, Z. Advances in 5-aminolevulinic acid priming to enhance plant tolerance to abiotic stress. *Int J Mol Sci***23**, 702 (2022).

18. Li, Y. *et al.* Isoleucine enhances plant resistance against *Botrytis cinerea* via jasmonate signaling pathway. *Front Plant Sci***12**, 628328 (2021).
19. Sadak, M. S. & Ramadan, A. A. E. M. Impact of melatonin and tryptophan on water stress tolerance in white lupine (*Lupinus termis* L.). *Physiology and Molecular Biology of Plants***27**, 469–481 (2021).
20. Lu, L. *et al.* Nuclear factor Y subunit GmNFYA competes with GmHDA13 for interaction with GmFVE to positively regulate salt tolerance in soybean. *Plant Biotechnol J***19**, 2362–2379 (2021).
21. Zhao, Q. *et al.* Na2CO3-responsive mechanisms in halophyte *Puccinellia tenuiflora* roots revealed by physiological and proteomic analyses. *Sci Rep***6**, 32717 (2016).
22. Wenkel, S. *et al.* CONSTANS and the CCAAT box binding complex share a functionally important domain and interact to regulate flowering of *Arabidopsis*. *Plant Cell***18**, 2971–2984 (2006).
23. Wang, Y. *et al.* Functional characterization of a sugar beet *BvbHLH93* transcription factor in salt stress tolerance. *Int J Mol Sci***22**, 3669 (2021).
24. Wang, J., Ye, Y., Xu, M., Feng, L. & Xu, L. A. Roles of the *SPL* gene family and miR156 in the salt stress responses of tamarisk (*Tamarix chinensis*). *BMC Plant Biol***19**, 370 (2019).
25. Shen, W. *et al.* Genomic and transcriptomic analyses of HD-Zip family transcription factors and their responses to abiotic stress in tea plant (*Camellia sinensis*). *Genomics***111**, 1142–1151 (2019).
26. Shah, W. H. *et al.* Understanding the integrated pathways and mechanisms of transporters, protein kinases, and transcription factors in plants under salt stress. *Int J Genomics***2021**, 5578727 (2021).
27. Amin, A. B. *et al.* Crassulacean acid metabolism abiotic stress-responsive transcription factors: a potential genetic engineering approach for improving crop tolerance to abiotic stress. *Front Plant Sci***10**, 129 (2019).
28. Yuan, G. *et al.* Biosystems design to accelerate C3-to-CAM progression. *BioDesign Research***2020**, 3686791 (2020).
29. Zhang, H. *et al.* Full-length transcriptome analysis of the halophyte *Nitraria sibirica* Pall. *Genes (Basel)***13**, 661 (2022).
30. Gish, L. A. & Clark, S. E. The RLK/Pelle family of kinases. *Plant Journal***66**, 117–127 (2011).
31. Orozco-Arias, S., Isaza, G. & Guyot, R. Retrotransposons in plant genomes: structure, identification, and classification through bioinformatics and machine learning. *Int J Mol Sci***20**, 3837 (2019).
32. Golda, M., Mótyán, J. A., Mahdi, M. & Tőzsér, J. Functional study of the retrotransposon-derived human PEG10 protease. *Int J Mol Sci***21**, 2424 (2020).
33. Sato, R. *et al.* NaCl–promoted respiration and cell division in halophilism of a halophyte, the common ice plant *Mesembryanthemum crystallinum* L. *Journal of the Faculty of Agriculture, Kyushu University***67**, *In press* (2022).
34. Chaudhry, U. K., Gökçe, Z. N. Ö. & Gökçe, A. F. The influence of salinity stress on plants and their molecular mechanisms. *Biology and life sciences forum***1**, x (2021).
35. Reyes-Pérez, J. J. *et al.* Evaluation of glycosyl-hydrolases, phosphatases, esterases and proteases as potential biomarker for NaCl-stress tolerance in *Solanum lycopersicum* L. Varieties. *Molecules***24**,

2488 (2019).

36. Winter, K., Ltittge, U., Winter, E. & Troughton, J. H. Seasonal shift from C3 photosynthesis to crassulacean acid metabolism in *Mesembryanthemum crystallinum* growing in its natural environment. *Oecologia (Berl.)***34**, 225–237 (1978).
37. Tran, D. Q., Konishi, A., Morokuma, M., Toyota, M. & Agarie, S. NaCl-stimulated ATP synthesis in mitochondria of a halophyte *Mesembryanthemum crystallinum* L. *Plant Prod Sci***23**, 129–135 (2020).
38. Agarie, S. *et al.* Potential of the common ice plant, *Mesembryanthemum crystallinum* as a new high-functional food as evaluated by polyol accumulation. *Plant Prod Sci***12**, 37–46 (2009).
39. Hunt, M. *et al.* REAPR: a universal tool for genome assembly evaluation. *Genome Biol***14**, R47 (2013).
40. Liu, Y., Schröder, J. & Schmidt, B. Musket: a multistage *k*-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics***29**, 308–315 (2013).
41. Simon, A. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (2010).
42. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics***27**, 764–770 (2011).
43. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference - free profiling of polyploid genomes. *Nat Commun***11**, 1432 (2020).
44. Pryszcz, L. P. & Gabaldón, T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res***44**, e113 (2016).
45. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: assessing genomic data quality and beyond. *Curr Protoc***1**, e323 (2021).
46. Nishimura, O., Hara, Y. & Kuraku, S. GVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics***33**, 3635–3637 (2017).
47. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res***32**, 20–25 (2004).
48. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A***117**, 9451–9457 (2020).
49. Abrusán, G., Grundmann, N., Demester, L. & Makalowski, W. TEclass - a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics***25**, 1329–1330 (2009).
50. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. TRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res***49**, 9077–9096 (2021).
51. Cognat, V. *et al.* PlantRNA, a database for tRNAs of photosynthetic eukaryotes. *Nucleic Acids Res***41**, 273–279 (2013).
52. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics***29**, 2933–2935 (2013).

53. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform***3**, 1–11 (2021).

54. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One***11**, e0163962 (2016).

55. Potter, S. C. *et al.* HMMER web server: 2018 update. *Nucleic Acids Res***46**, W200–W204 (2018).

56. Sherman, B. T. *et al.* DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res***50**, W216–W221 (2022).

Figures

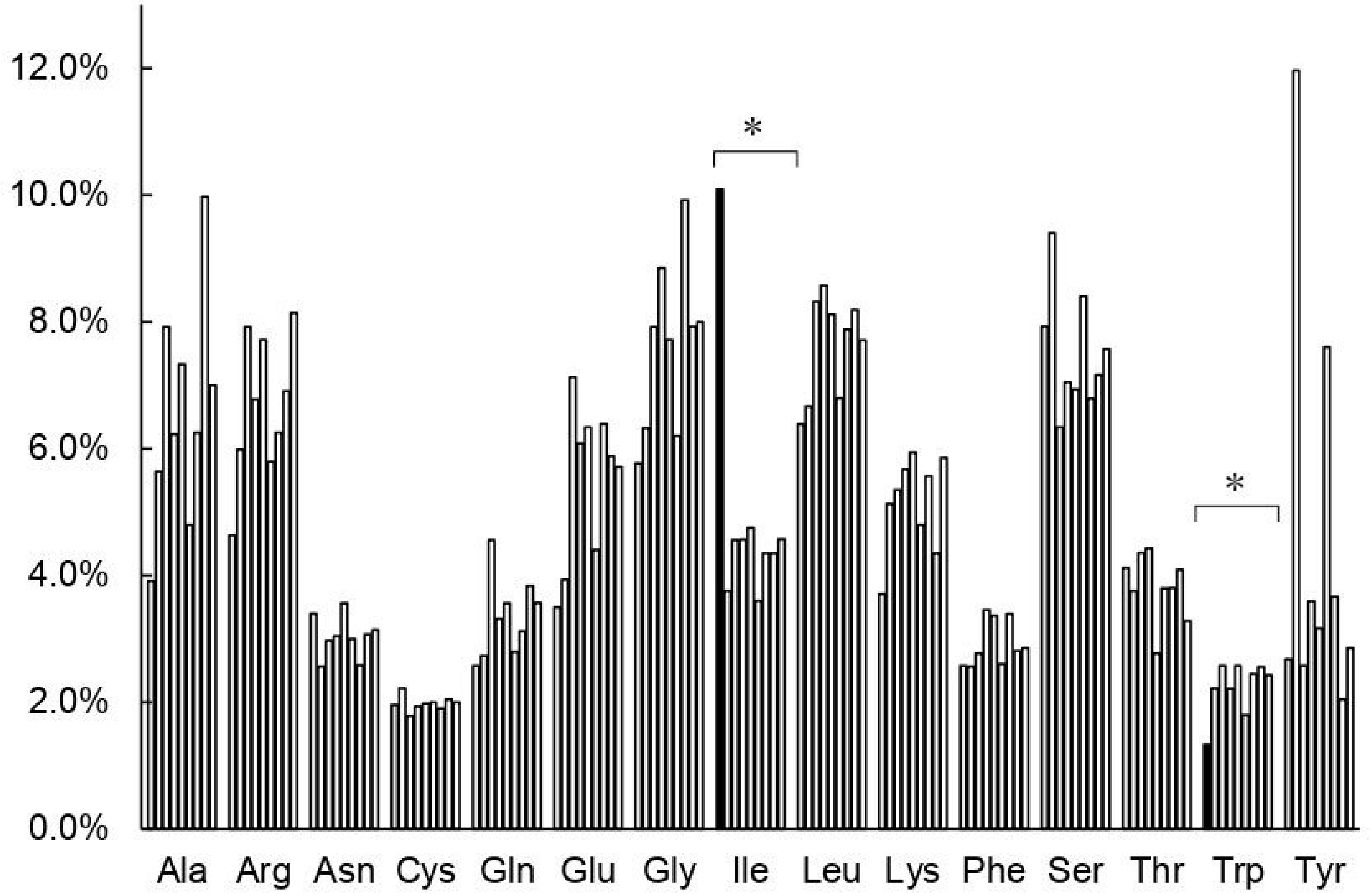

Figure 1

Comparison of the percentage of tRNAs in 9 plant species including ice plant. tRNAs significant differently abundant from the other 8 species by Smirnov-Grabs outlier test, are shown in black, and the other tRNAs are shown in gray. Bars indicate ice plant, *Arabidopsis*, rice, tomato, poplar, horseradish, potato, grape, and soybean from the left of each series. Asterisks indicate statistical significance: * $P < 0.05$, n = 9.

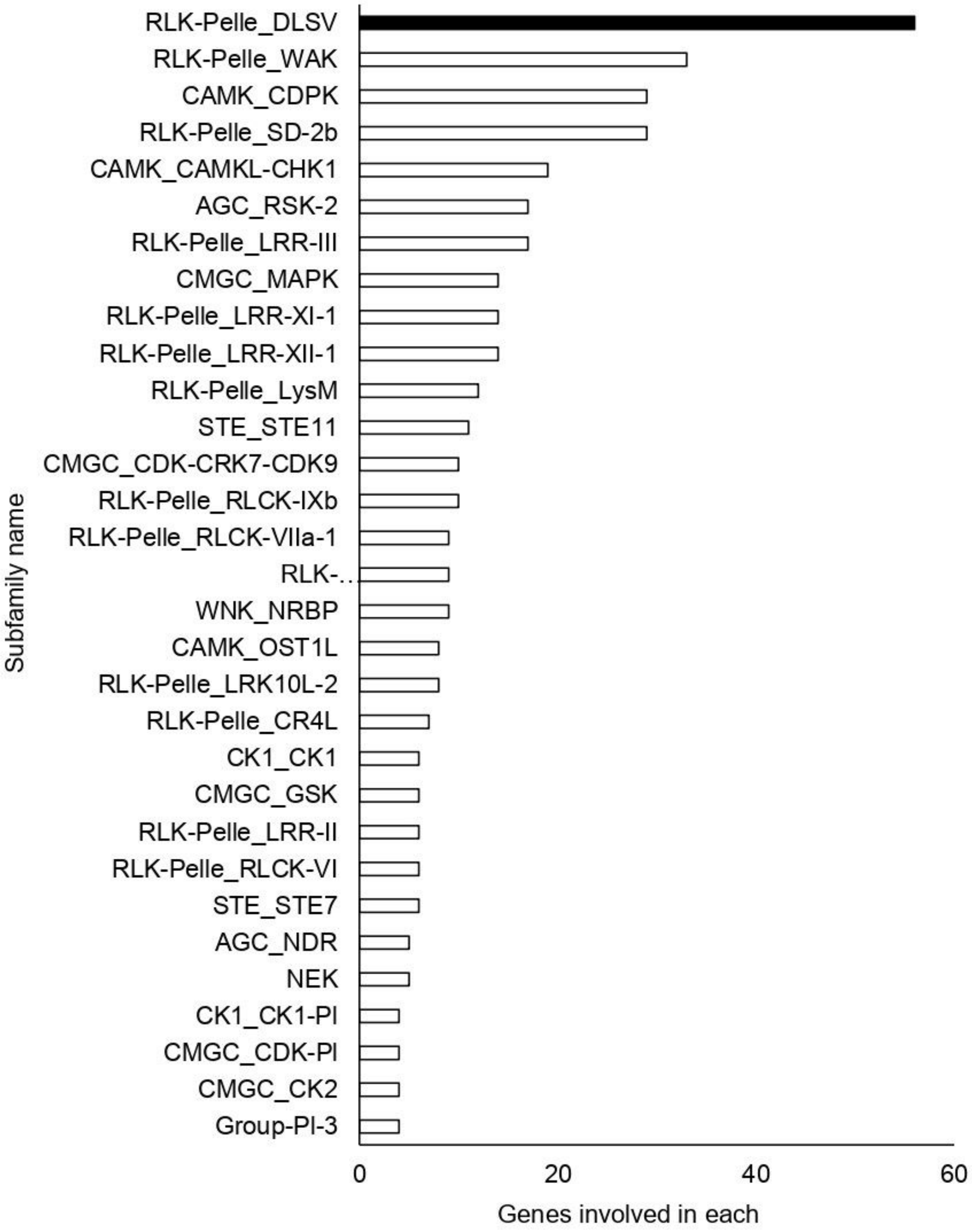

Figure 2

Top 30 Pkinase subfamilies classified in descending order of the number of genes included in them. The family with the highest number of genes is shown in black, and the other families are shown in white.

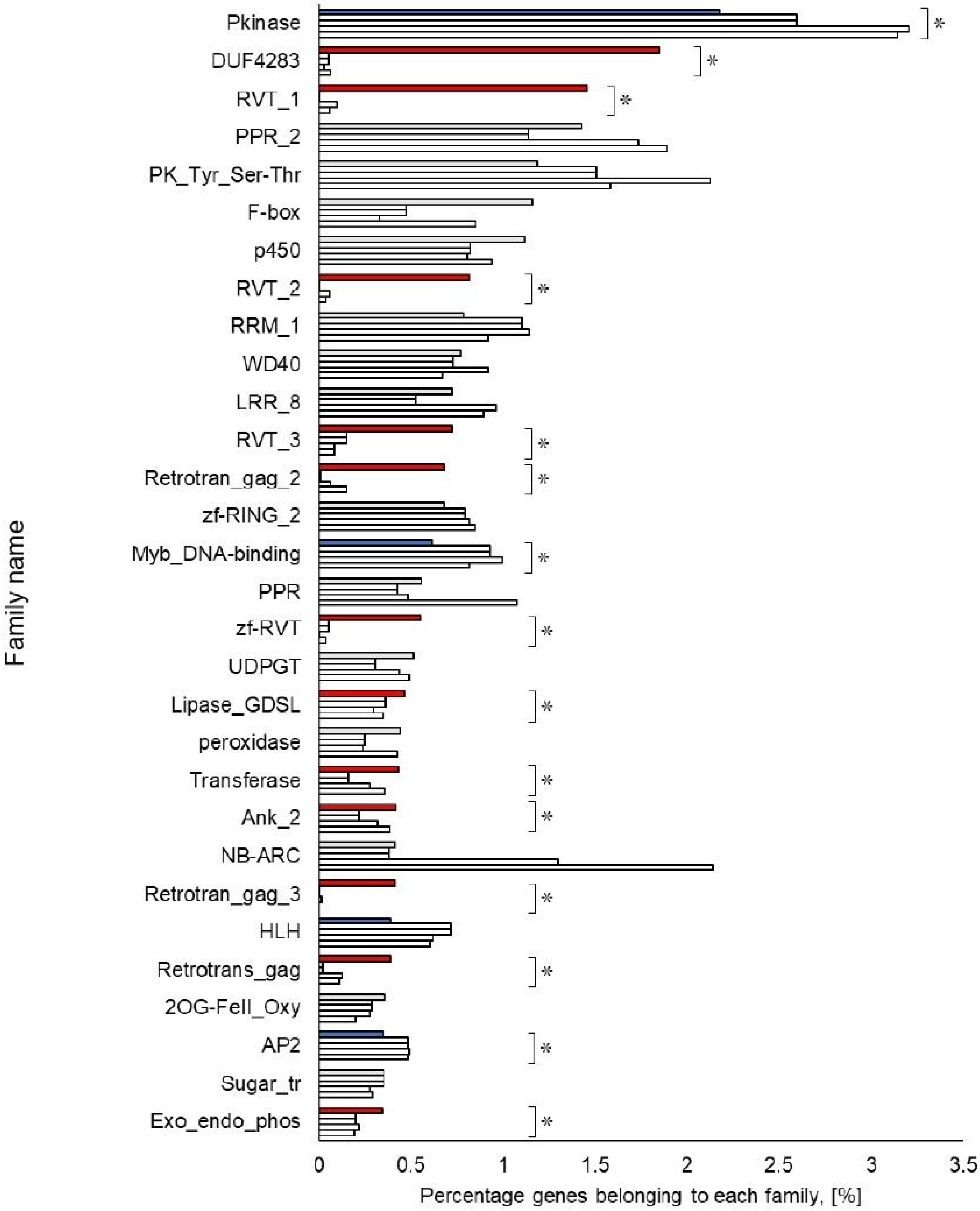

Figure 3

Top 30 gene families obtained from amino acid sequences detected in ice plant four other plant species. The top row for each family shows ice plant, *Arabidopsis*, rice, maize, and poplar. The independence of the proportion of genes belonging to a family in the ice plant is displayed using the Smirnov-Grubbs rejection test. Asterisks (*) indicate statistical significance: *P* < 0.05, n=5. Independence is shown in red if the proportion is independently high in ice plant, in blue if it is low, and in gray if there is no difference.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf
- SupplementaryDatasetS1.xlsx
- SupplementaryDatasetS2.xlsx
- SupplementaryDatasetS3.xlsx