frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 30 June 2022
doi: 10.3389/fpls.2022.923069

Taxonomically Restricted Genes Are Associated With Responses to Biotic and Abiotic Stresses in Sugarcane (*Saccharum* spp.)

Cláudio Benício Cardoso-Silva[1,2†], Alexandre Hild Aono[1], Melina Cristina Mancini[1], Danilo Augusto Sforça[1], Carla Cristina da Silva[1,3], Luciana Rossini Pinto[4], Keith L. Adams[2] and Anete Pereira de Souza[1,5]*

[1]Center of Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Campinas, Brazil, [2]Department of Botany, University of British Columbia, Vancouver, BC, Canada, [3]Agronomy Department, Federal University of Viçosa (UFV), Viçosa, Brazil, [4]Sugarcane Research Advanced Centre, Agronomic Institute of Campinas (IAC/APTA), Ribeirão Preto, Brazil, [5]Institute of Biology, University of Campinas (UNICAMP), Campinas, Brazil

Orphan genes (OGs) are protein-coding genes that are restricted to particular clades or species and lack homology with genes from other organisms, making their biological functions difficult to predict. OGs can rapidly originate and become functional; consequently, they may support rapid adaptation to environmental changes. Extensive spread of mobile elements and whole-genome duplication occurred in the *Saccharum* group, which may have contributed to the origin and diversification of OGs in the sugarcane genome. Here, we identified and characterized OGs in sugarcane, examined their expression profiles across tissues and genotypes, and investigated their regulation under varying conditions. We identified 319 OGs in the *Saccharum spontaneum* genome without detected homology to protein-coding genes in green plants, except those belonging to Saccharinae. Transcriptomic analysis revealed 288 sugarcane OGs with detectable expression levels in at least one tissue or genotype. We observed similar expression patterns of OGs in sugarcane genotypes originating from the closest geographical locations. We also observed tissue-specific expression of some OGs, possibly indicating a complex regulatory process for maintaining diverse functional activity of these genes across sugarcane tissues and genotypes. Sixty-six OGs were differentially expressed under stress conditions, especially cold and osmotic stresses. Gene co-expression network and functional enrichment analyses suggested that sugarcane OGs are involved in several biological mechanisms, including stimulus response and defence mechanisms. These findings provide a valuable genomic resource for sugarcane researchers, especially those interested in selecting stress-responsive genes.

Keywords: orphan genes, sugarcane hybrid, stress condition, RNA-Seq, gene expression

OPEN ACCESS

Edited by:
Andrew H. Paterson, University of Georgia, United States

Reviewed by:
Igor Fesenko, Institute of Bioorganic Chemistry (RAS), Russia
Youxiong Que, Fujian Agriculture and Forestry University, China
Rasappa Viswanathan, Indian Council of Agricultural Research (ICAR), India

***Correspondence:**
Anete Pereira de Souza
anete@unicamp.br

†Present address:
Cláudio Benício Cardoso-Silva, Laboratório de Química e Função de Proteínas e Peptídeos, Universidade Estadual do Norte Fluminense, Rio de Janeiro, Brazil

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *18 April 2022*
Accepted: *13 June 2022*
Published: *30 June 2022*

Citation:
Cardoso-Silva CB, Aono AH, Mancini MC, Sforça DA, da Silva CC, Pinto LR, Adams KL and de Souza AP (2022) Taxonomically Restricted Genes Are Associated With Responses to Biotic and Abiotic Stresses in Sugarcane (Saccharum spp.). Front. Plant Sci. 13:923069. doi: 10.3389/fpls.2022.923069

INTRODUCTION

Recent advances in sugarcane genomics have created opportunities to systematically reveal the evolutionary history and diversification of the *Saccharum* group. However, the complexity of the sugarcane genome, mainly due to its size, ploidy level, and large number of mobile elements (Thirugnanasambandam et al., 2018), has hindered advances in the genomics of this important crop species. Despite the economic importance of sugarcane due to its use as a source of sugar, biofuel, and fibre, its reference genomes, including a chromosome-level *Saccharum spontaneum* genome (Zhang et al., 2018), a monoploid genome from the R570 variety (Garsmeur et al., 2018), and an SP80-3280 hybrid genome (Souza et al., 2019), have been only recently reported.

Two events of whole-genome duplication (WGD) are thought to have occurred during the evolution of the *Saccharum* group (Ming et al., 1998; Paterson et al., 2012). WGD is a major mechanism responsible for species diversification and adaptation (Soltis et al., 2009; Renny-Byfield and Wendel, 2014). These recent events of polyploidization occurring within the Saccharinae group provide an opportunity to investigate the fate of duplicated genes. Genome duplication initially results in gene duplication and gene redundancy. After duplication, some gene copies preserve their original function, while most of them are eliminated through negative selection (Tautz and Domazet-Lošo, 2011). However, some copies under positive selection, after sequence diversification, may acquire a new biological function (Van de Peer et al., 2009). Divergence of pre-existing genes is one of the mechanisms underlying the emergence of new genes (Tautz and Domazet-Lošo, 2011). An alternative origin has been proposed: new genes originate from a non-coding sequence (Singh and Syrkin Wurtele, 2020; Vakirlis et al., 2020).

Taxonomically restricted, lineage-specific or orphan genes (OGs), which have no homology to genes in other taxa, may contribute to evolutionary novelties and might be responsible for some lineage-specific trait origins (Wilson et al., 2005; Khalturin et al., 2009; Tautz and Domazet-Lošo, 2011). Even though we have not given sufficient attention to these genes, comparative genomic studies have estimated that OGs constitute at least 1% of the total genes in a genome, depending on the alignment rate and taxonomic level considered (Khalturin et al., 2009; Arendsee et al., 2014; Prabh and Rödelsperger, 2016). Several studies have been carried out to characterize OGs in plants at the species level: Arabidopsis (Li and Wurtele, 2014), sweet orange (Xu et al., 2015), rice (Guo et al., 2007), moso bamboo (Zhang et al., 2022), and at the family level: Brassicaceae (Donoghue et al., 2011) and Poaceae (Campbell et al., 2007). However, there is limited information about the function of most of these OGs, as they lack recognizable domains and functional motifs.

OGs are known to play a role in primary metabolism and response to environmental changes in plants. By establishing a gene-editing system, Jiang et al. (2020) revealed that an orphan gene (*BrOGs*) in *Brassica napus* plays a vital role in soluble sugar metabolism. In *Arabidopsis*, a functional analysis of the well-studied orphan gene *QQS* (*qua quine starch*) indicated that it could act in regulation of nitrogen allocation, affecting the protein content (O'Conner et al., 2018). Additionally, there are several works showing that OGs are regulated in response to biotic and abiotic stresses (Beike et al., 2014; Giarola et al., 2014; Khraiwesh et al., 2015; Schlötterer, 2015; Kaur et al., 2017). For example, an OG named *TaFROG* enhanced wheat resistance to Fusarium head blight (Perochon et al., 2015) and an OG in *Vigna unguiculata* (*UP12_8740*) increased plant tolerance to osmotic stresses and soil drought (Li et al., 2019). A *Physcomitrium patens* OG (*PpARDT*) was functionally characterized, and knockout mutant displayed reduced drought tolerance (Dong et al., 2022).

Despite the biological relevance of these taxonomically restricted genes, no previous reports described their occurrence and expression profile in the *Saccharum* complex. To advance our knowledge about OGs in sugarcane, a comparative genomic approach is needed for their identification, followed by regulatory inference based on gene expression analysis. In this study, we identified and characterized sugarcane OGs and their expression patterns across tissues and genotypes. Additionally, we analysed expression data from different conditions to identify those under which these genes are positively or negatively regulated in sugarcane.

MATERIALS AND METHODS

Orphan Gene Identification

A phylostratigraphic approach based on a sequence homology search was used to identify OGs in the sugarcane genome (Domazet-Lošo et al., 2007; McLysaght and Hurst, 2016). These analyses rely on sequence alignments to detect genes that lack homology in a focal species in comparison with a target clade. For this analysis, we used the gene model from *S. spontaneum* (Zhang et al., 2018) as a reference. First, the protein and coding DNA sequence (CDS) files containing the set of sugarcane genes were filtered using the CD-HIT package v.4.8.1 (Fu et al., 2012); a similarity threshold of 90% was applied for both the CD-HIT and CD-HIT-EST algorithms, which were employed for the protein and CDS files, respectively. This step was performed to remove redundancies in the dataset once all the homologous genes and duplications were included in the annotated sugarcane genome. Subsequently, a series of local alignments using both the sugarcane proteome and CDSs were performed to remove genes with homology in other species. First, to reduce the subset of candidate genes, we filtered out all sugarcane genes with detected homology to genes annotated in 13 angiosperm species including seven Poaceae species (*Arabidopsis thaliana* TAIR10, *Brachypodium distachyon* v3.1, *Citrus sinensis* v3.1, *Eucalyptus grandis* v2.0, *Miscanthus sinensis* 7.1, *Oryza sativa* 7.0, *Phaseolus vulgaris* v2.1, *Panicum virgatum* v4.1, *Setaria italica* v2.2, *Solanum lycopersicum* ITAG3.2, *Sorghum bicolor* v3.1.1, *Oropetium thomaeum* v1.0, and *Zea mays* 284 v6). The annotated sequences were downloaded from Phytozome v.13 (Goodstein et al., 2011) and converted into a database. The remaining subset of sugarcane genes was aligned to

non-redundant proteins (NR) and nucleotides (NT) from the National Center for Biotechnology Information (NCBI) database, and genes without homology in previous filtering steps were discarded. In all filtering steps, we used BLASTp and BLASTn to detect homology at the protein and nucleotide levels, respectively. All results were generated using a permissive E-value cut-off ≤10–6, allowing more relationships to be detected and increasing the chance of selecting a real OG. To discard the hypothesis that predicted OGs were missing from the genome annotation, we mapped each OG back on the chromosomes of seven species downloaded from Phytozome v.13 (*A. thaliana*, *Panicum hallii*, *S. italica*, *Z. mays*, *S. bicolor*, *S. spontaneum*, and *M. sinensis*) using sim4 software, which employs a splice-aware alignment method (Florea et al., 1998).

Orphan Gene Characteristic Features and Sequence Homology

The FASTA files containing sugarcane chromosome information and gff3 files were used for manual curation of the OGs. The position of each OG exon was used as a starting point to check intron/exon boundaries as well as the presence of start and stop codons using Artemis software v.18.0 (Carver et al., 2011). To characterize the physical and chemical properties of sugarcane genes (OGs and non-OGs), we calculated protein parameters [protein length, molecular weight, the instability index, hydrophobicity, the isoelectric point, and the grand average of hydropathicity (GRAVY)] using ProtParam tools implemented in the Bio.SeqUtils package, a Biopython module (Cock et al., 2009). To assess the protein-coding potential of these genes, we estimated the probability of each OG being a coding RNA using Coding Potential Calculator (CPC2; Kang et al., 2017). Additionally, all predicted OGs were aligned to the non-coding RNA databases derived from Rfam, Tair, Ensemble long non-coding RNAs (lncRNAs), CANTATA (Szcześniak et al., 2015), and GreeNC (Paytuví Gallart et al., 2015) using nhmmer with an E-value parameter ≤0.001 (Wheeler and Eddy, 2013). We checked for transposable element (TE) insertion in the OG sequences by using a reference collection of transposons from the Repbase database (Bao et al., 2015) using CENSOR (Jurka et al., 1996). To search for homologues within the Saccharinae subtribe, we aligned predicted proteins of OGs to annotated proteins from the draft genomes of *Saccharum* hybrids SP803280 (Souza et al., 2019) and R570 (Garsmeur et al., 2018) and complete genomes from *M. sinensis* (Mitros et al., 2020) and *S. spontaneum* (Zhang et al., 2018) using BLASTp with an E-value $\leq 10^{-6}$.

RNA-Seq Experimental Data: Retrieval and Pre-processing

An extensive search for papers reporting RNA-Seq data in sugarcane was performed, followed by a search of the NCBI Sequence Read Archive (SRA) repository (**Supplementary Table 1**). The selected RNA-Seq samples were retrieved using the 'fastq-dump' program from the SRA toolkit (version 8.22), and SRA files were converted to fastq-format files. The raw reads were subjected to quality control using Trimmomatic v0.36 (Bolger et al., 2014) to remove adapter and low-quality sequences. Three reference transcriptomes, including two full-length transcriptomes, were also selected to confirm the selected OGs being transcribed. In the first set of IsoSeq data, RNA samples were extracted from the top and bottom internodes of 22 genotypes (Hoang et al., 2017), and in the second set, RNA samples were obtained from leaves of a commercial sugarcane variety from Thailand (Piriyapongsa et al., 2018). The third transcriptome, which was *de novo* assembled from short reads, was extracted from the leaves of six sugarcane hybrids (Cardoso-Silva et al., 2014). A local alignment using the BLASTn program was performed with an e-value cut-off $\leq 1 \times 10^{-6}$ to infer the homology of putative OGs to sugarcane transcripts.

Orphan Gene Expression Profile and Differential Expression

RNA-Seq libraries were constructed for two purposes: (i) to unveil the expression patterns of OGs across sugarcane tissues and genotypes and (ii) to identify DE OGs, especially under stress conditions. The expression level of each gene was estimated by mapping the transcriptomes against the whole gene set of sugarcane using Salmon (Patro et al., 2017). The expression level of a given gene was calculated by the log transformation method implemented in Salmon [transcripts per million (TPM)], which represents the relative abundance of a transcript among a population of transcripts. Heatmaps representing the expression levels of OGs in experiments testing for differential gene expression and evaluating expression across tissues/genotypes were created using the 'superheat' R package (Barter and Yu, 2018).

To investigate whether OGs were DE, we designed RNA-Seq experiments including biotic and abiotic stresses, developmental stages, and sucrose accumulation. Cleaned reads from each library originating from experiments with biological replicates were mapped to the complete set of sugarcane genes (CDS FASTA format) using Salmon v.0.12.0 to quantify transcript abundance (Patro et al., 2017). The DESeq2 package v.3.9 (Love et al., 2014) was used to predict DEGs in each experiment using the raw read counts as input data. In cases where the same sample was sequenced in multiple runs, the technical replicates were collapsed before starting the DEG analysis. To minimize quantification biases, genes with fewer than 10 reads mapped per sample were filtered out before the gene expression analysis. The DEGs were estimated assuming a negative binomial distribution for each gene, applying a function that estimates the size factor and reducing bias caused by library size (normalization by the median ratio; gene count divided by the sample size). A value of $p < 0.05$ and an absolute log2 fold change ≥2 were used as thresholds for determining whether genes were apparently DE. For each predicted OG, hypothesis testing was performed, in which the null hypothesis was no differences between the control and treatment groups, thus supporting the assumption that any difference in gene expression occurred merely by chance.

Orphan Gene Co-expression Network and Functional Annotation

The expression matrix of the 218 samples across sugarcane tissues and genotypes was used to build a co-expression network with the 'WGCNA' R package (Langfelder and Horvath, 2008). A weighted adjacency matrix was constructed using pairwise Pearson's correlation coefficient measures and an estimated power threshold for scale-free independence ($R^2 > 0.8$ and largest mean connectivity). Subsequently, the calculated matrix was converted to a topological overlap matrix (TOM), which evaluates gene pair correlations and the degree of agreement with other genes in the matrix (Yip and Horvath, 2007). After that, we inferred network functional modules by employing average linkage hierarchical clustering in accordance with the TOM-based dissimilarity measure. We used a soft threshold power of 7 (R^2 of 0.81 and mean connectivity of 164) to calculate the TOM. Clusters were defined according to a hierarchical dendrogram using adaptive branch pruning as implemented in the 'dynamicTreeCut' R package (Langfelder et al., 2007). We conducted functional enrichment analysis of the modules containing OGs based on Gene Ontology terms. Then, we assumed a guilty-by-association approach to obtain some insight into OG functionality. Additionally, we checked for OG similarity with known protein domains using hmmscan from the HMMER3 suite (Finn et al., 2011), aligning the domains to the Pfam v35 database (Mistry et al., 2020).

RESULTS

Identification and Characterization of Sugarcane Orphan Genes

After removing redundancies in the sugarcane gene model, which contained 83,826 genes from *S. spontaneum* (Zhang et al., 2018), we obtained a total of 51,675 NR protein-coding genes. These sugarcane genes were aligned to the proteomes of 13 angiosperm species, including seven Poaceae species, represented by 435,957 proteins. This alignment returned 1,536 sugarcane genes with no homology with any protein represented. Next, these subsets of 'no-hit' genes were aligned to the NR protein database, and 442 genes with no homology were found. Finally, these remaining sets were aligned to the NT database. As a result, a total of 335 genes were identified as sugarcane OGs due to their lack of homology to other genes.

Homology searches may fail to detect homologues in other species, resulting in spurious OG prediction. To minimize this effect, we did not rely only on homology searches but also mapped OG CDSs onto each chromosome of seven grass genomes (*S. spontaneum*, *M. sinensis*, *S. bicolor*, *P. hallii*, *Z. mays*, *S. italica*, and *O. sativa*) and the genome of *A. thaliana*. Although OGs were not annotated as complete genes, except in the *Saccharum* group, we found vestiges of exons of these genes in all grass chromosomes. However, we did not detect OG vestiges in *A. thaliana* chromosomes (**Supplementary Table 2**).

To better understand the distribution of these putative OGs across grass genomes, we assessed chromosome regions in which dispersed fragments of the OGs aligned with at least 10% of their length. Intriguingly, the closer the phylogenetic relationship with *Saccharum* was, the higher the number of OG fragments observed in grass genomes (**Figure 1**; **Supplementary Table 2**). The number of OG fragments ranged from 114 in the *O. sativa* genome to 91,379 in the *M. sinensis* genome. If we consider the *S. spontaneum* genome, the number of OG fragments is even larger. Curiously, there are 25 times more OG fragments in *S. spontaneum* and *M. sinensis* chromosomes than in the sorghum genome, which is the closest relative of these Saccharinae species.

To verify whether OGs are evolutionarily conserved within the Saccharinae subtribe, we searched for OGs homologs in the *Saccharum* spp., including the annotated genome of *M. sinensis*. We found 201 OGs with similarity to other sequences predicted to be protein-coding genes in Saccharinae genomes. Most homologs were present in the *Saccharum* hybrid genome; however, we also found 39 homologs genes in the *M. sinensis* genome. Furthermore, 127 OGs have duplicate copies in *S. spontaneum*, annotated as alleles or paralogues (**Supplementary Figure 1**; **Supplementary Table 3**). Additionally, because many OGs were not predicted as coding genes in other *Saccharum* genomes, we aligned them to non-coding RNA sequences. However, we did not detect similarity of any OGs with ncRNAs deposited in public databases, including lncRNAs.

Using a vector machine-based classifier named CPC2, 152 OGs were classified as ncRNAs, and 167 OGs were classified as coding RNAs (**Supplementary Figure 2**; **Supplementary Table 4**). A total of 89 OGs were predicted as protein-coding genes with a high probability ($>=0.9$), while 47 others were classified as lncRNAs. However, short OGs were more likely to be classified as ncRNAs ($R^2 = 0.73$, value of $p < 2.2e^{-16}$; **Supplementary Figure 3**). A search for protein domains within OG sequences revealed that most OGs do not show similarity with any functional domains deposited in the Pfam database. Based on these searches, we detected only partial local alignment for 30 OGs (BLAST searches produced no significant alignments to domains; E-value $\leq$1E-3), of which nine were predicted to be domains of unknown function (**Supplementary Table 4**).

Physical and chemical analyses revealed that OGs and non-OGs were significantly different in all parameters except GC content (**Supplementary Figure 4**; **Supplementary Table 4**). OGs are shorter than non-OGs, with average protein lengths of 136 and 450, respectively. The number of exons in OGs varied from 1 to 21 ($\bar{x} = 3.54$; $\sigma = 2.37$), and 57.9% had three exons or fewer (**Supplementary Table 4**). The average GC contents of orphan and non-OGs were almost identical, at 56.7 and 56.5%, respectively. Most sugarcane genes had a negative GRAVY (grand average of hydropathy) value, i.e., ~81% of the OGs and ~80% of the non-OGs, which supports the protein being hydrophilic.

TE Fragments in the OGs

The large number of gene fragments found in the Saccharinae genomes may indicate that some of these genes originated from TE duplication. We aligned all the predicted OGs to TE sequences to test the hypothesis that some of the OGs are TEs and to verify whether some of the OGs were derived

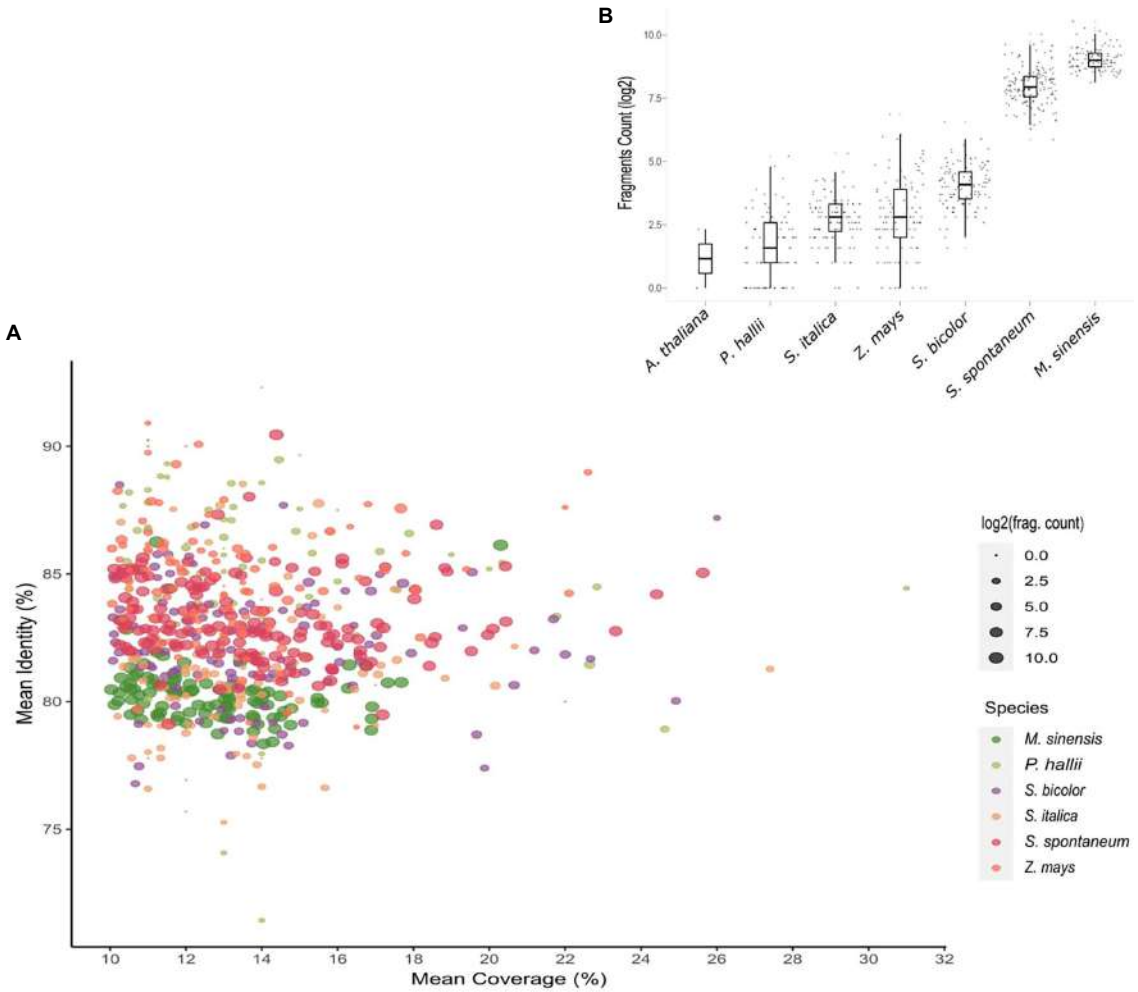

FIGURE 1 | Evidence of orphan gene (OG) vestiges within grass genomes. **(A)** Scatterplot showing the average percentage identity and coverage of each OG in six grass species. Each dot represents an individual OG. The dot sizes represent the number of fragments of each sugarcane OG in other grass species. The number of fragments of each OG in the focal species is also shown in the bar plot **(B)**.

from TE insertion (**Supplementary Table 5**). This search was motivated by previous observations suggesting that 51% of the OGs in rice are derived from TEs (Jin et al., 2019). To investigate this hypothesis more deeply and shed light on the origin of these putative lineage-specific genes, we performed an alignment of the initially selected set of genes (335 OGs) against the Repbase TE library.[1] A total of 153 putative OGs aligned to TEs with significant hits (E-value ≤1E-10). For a subset of these genes (16 OGs), at least 70% of the sequence aligned to TEs. We assumed that these genes were putative TEs, and we did not consider them to be sugarcane OGs. In the remaining subset (319 OGs), some genes had traces of TE insertion into the coding region, albeit with poor alignment. To better understand this result, we estimated the fractions of both OGs and non-OGs in sugarcane with similarity to TEs. Most sugarcane genes had traces of TEs in their coding region, and ~55% of the OGs and ~75% of the non-OGs had at least 10% of their sequence aligning to TEs (**Supplementary Figure 5**).

Evidence of Orphan Gene Expression Across Sugarcane Tissues and Genotypes

We searched for evidence that the 319 OGs were being transcribed across several sugarcane tissues and genotypes by aligning them against reference transcriptomes and RNA-Seq libraries. We selected two representative sugarcane IsoSeq datasets (Hoang et al., 2017; Piriyapongsa et al., 2018), a collection of transcripts from six sugarcane varieties (Cardoso-Silva et al., 2014), and 218 RNA-Seq samples from sugarcane hybrids, *Saccharum officinarum,* and *S. spontaneum* (**Supplementary Table 1**). Evidence of transcription was detected in 89.34% of the OGs (TPM value ≥1), which were expressed in at least one transcriptome experiment or tissue (**Supplementary Table 6**). Almost one-third of the OGs had an expression level considered low (1 ≤ TPM ≤ 10), while only 6% of them had a value greater than 100 TPM. In fact, when

[1]http://www.girinst.org/repbase

we compared the expression levels of OGs and non-OGs in four sugarcane tissues (**Supplementary Figure 6**), we found that OGs had proportionally lower expression levels. However, these differences were almost imperceptible in the meristem tissue (bud).

The expression profile of OGs across tissues and genotypes revealed a clear pattern of sample clustering. Overall, OGs showed similar expression patterns among genotypes when we compared the same tissue (**Figure 2** and **Supplementary Table 6**). Although the expression matrix combined tissues and genotypes, we also observed clustering by genotype to the same degree. For example, samples originating from *S. spontaneum* (Krakatau, IN84_58, and SES205A) were clustered together and separated from those originating from *S. officinarum* (BadilaDeJava, CriollaRayada, and WhiteTransparent) and *Saccharum* hybrids. Notably, samples originating from *Saccharum* hybrids and *S. officinarum* had more similar expression patterns than those originating from *S. spontaneum*. In particular, the expression pattern of OGs in the internodes of *S. spontaneum* was noticeably different from that in *S. officinarum* and hybrids. Furthermore, some OGs presented contrasting expression patterns between samples from *S. officinarum* and *S. spontaneum*. Specifically, a subset of 42 OGs had an average expression level that was three times higher in *S. officinarum* than in *S. spontaneum*. In contrast, the expression level of 35 OGs was higher in *S. spontaneum*. Interestingly, genotypes originating from the same geographical location tended to be clustered together because they had more similar expression profiles. In particular, this pattern was observed in the expression levels of OGs in internode 1 of French hybrids (F36.819 and R570) and the roots of Brazilian hybrids (RB855536, RB855113, and RB867515).

Interestingly, a few OGs seemed to have tissue-specific regulation, as observed for some genes only expressed in roots (Sspon.06G0001310 and Sspon.06G0024430), internode 1 (Sspon.08G0030340), internodes 3 and 8 (Sspon.05G0032460 and Sspon.08G0007980), and all internodes (Sspon.06G.0027080).

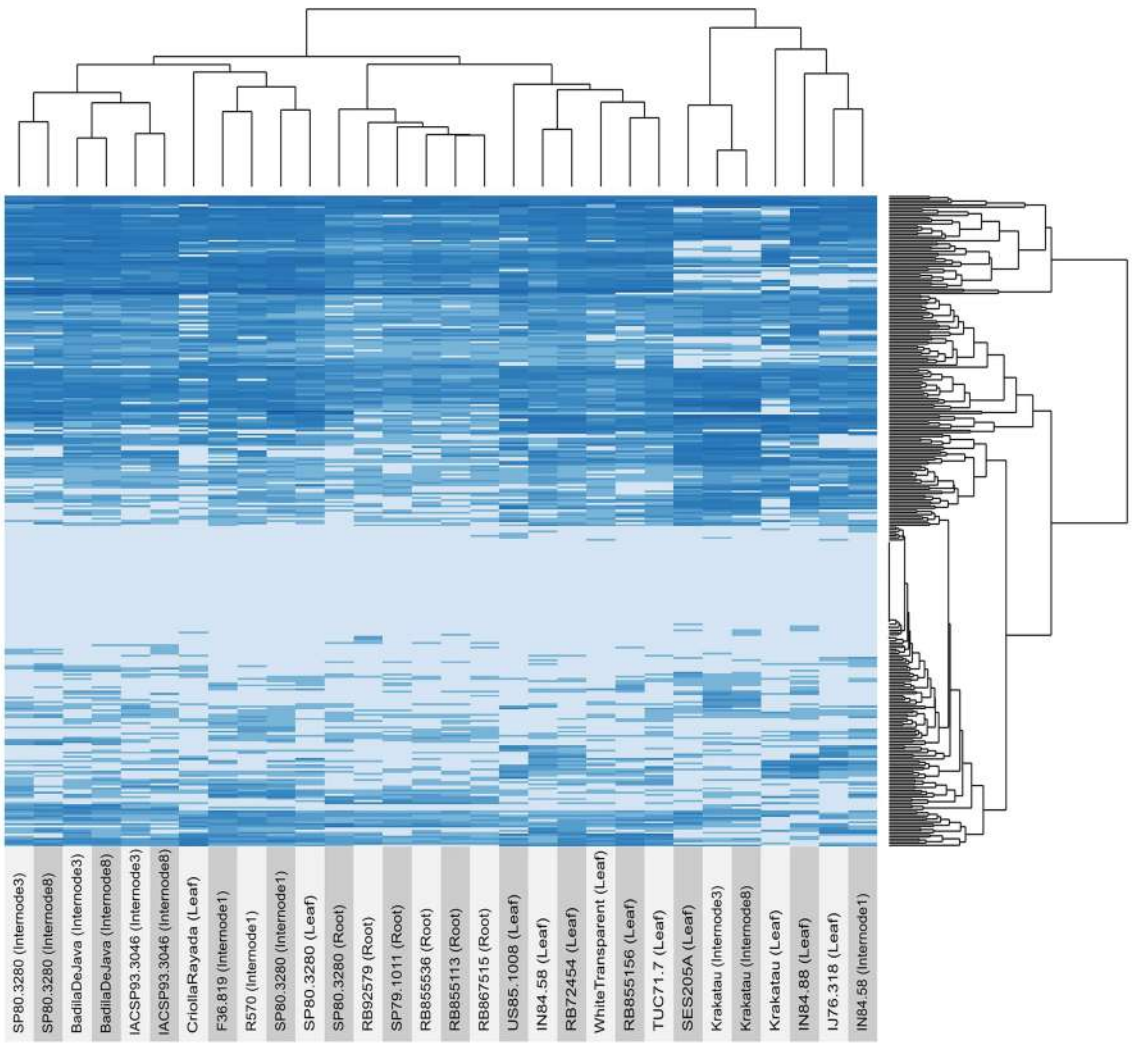

FIGURE 2 | Expression profiles of orphan genes (OGs) in several sugarcane tissues and genotypes. The expression of each gene (TPM) was estimated using a pseudoalignment method implemented in Salmon software.

Orphan Genes Are Differentially Expressed Under Stress Conditions

We performed eight RNA-Seq experiments representing a variety of conditions to test whether OGs change their expression levels (**Table 1**). After filtering the raw data, we selected more than 13 billion high-quality reads to perform gene expression analysis.

We observed at least one DE OG in five of the eight RNA-Seq experiments (**Supplementary Table 7**). We did not detect DE OGs in the RNA-Seq experiments related to developmental stage, sucrose accumulation, and plant infection with *Sporisorium scitamineum*. Generally, more genes were DE in the experiments related to abiotic stress than in those related to biotic stress. For example, we estimated that 6,440 genes were DE under cold stress (in genotypes Guitang08-1,180 and ROC22), while only 1,548 genes were DE after infection with yellow canopy syndrome, and 2,612 genes were DE in plants infected with smut disease. Overall, we identified 66 OGs that were DE under at least one type of stress ($\log_2$FoldChange ≥ 2; padj < 0.05). Most of the genes were regulated by abiotic stresses, while only 4 OGs were regulated by biotic stresses. We detected only two OGs (Sspon.02G0052680-1C and Sspon.06G0009900-2C) that were DE in both the osmotic stress and cold stress experiments, and one gene (Sspon.04G0024550-1B) showed DE in both the cold stress and low-nitrogen experiments.

In the cold stress experiment, we identified 8,921 DE genes in the hybrid Guitang08-1,180 (5,644 upregulated and 3,277 downregulated) and 8,913 DE genes in the hybrid ROC22 (5,596 upregulated and 3,317 downregulated), and 72% of these genes were DE in both genotypes. A total of 50 OGs were determined to be DE in this experiment, with 33 OGs in Guitang08-1,180 (12 upregulated and 21 downregulated) and 36 OGs in the ROC22 hybrid (12 upregulated and 24 downregulated), while 18 OGs were DE in both genotypes (**Figure 3** and **Supplementary Table 7**). An OG that was upregulated in both genotypes (Sspon.06G00100300) has three conserved copies annotated in the *S. spontaneum* genome (**Supplementary Figure 7**).

In the osmotic stress experiment conducted on leaves and root samples from *S. officinarum*, a total of 4,207 genes were DE in the leaves (1,815 upregulated and 2,392 downregulated), while 4,222 genes were DE in the root samples (1,707 upregulated and 2,515 downregulated). Of these genes, 13 OGs were identified as DE, nine in the leaf samples and six in the root samples. Two OGs, Sspon.01G0060030-1D and Sspon.05G0013120-1A, were detected as DE in both the root and leaf tissues (**Figure 4** and **Supplementary Table 7**).

In the sugarcane plants exposed to low-nitrogen conditions, most of the DE genes were detected in the leaf samples (4,524 genes in the Badila variety and 2,345 genes in the ROC22 hybrid), while in the roots, 894 and 726 genes were estimated to be DE in Badila and ROC22, respectively. In the root samples, the number of genes upregulated was three times greater than the number of genes downregulated. This pattern was observed in both genotypes. We did not detect DE OGs in the roots of either genotype; however, in the leaves, we detected six DE genes in Badila (four upregulated and two downregulated) and four DE genes in the ROC22 hybrid (one upregulated and three downregulated; **Figure 5** and **Supplementary Table 7**).

Co-expression Network and Modules With OGs

The co-expression network was built with the full set of sugarcane genes, including 288 OGs with estimated expression values. The sugarcane genes were distributed among 153 modules, of which 78 had at least one OG. More than one-third of the OGs (120 of 319) were assigned to seven modules. An enrichment analysis of these modules containing OGs (Fisher's exact test; value of $p \leq 0.05$) suggested that the genes are associated with several biological processes (**Supplementary Table 8**). Overall, the most frequent Gene Ontology terms observed in modules containing OGs included those associated with responses to several stimuli and defence mechanisms (**Supplementary Figure 8**). We identified a module with a set of co-expressed genes containing 64 OGs, which were enriched in Gene Ontology function terms related to transport (GO:0006810), response to starvation (GO:0042594), and response to external stimulus (GO:0009605). We also found modules enriched with genes associated with protein modification

TABLE 1 | RNA-Seq experiments performed for sugarcane gene expression analysis.

Accession	Condition	Cultivar/Species	Tissue	Rep[1]	References
PRJNA474042	yellow canopy syndrome	Hybrid	leaf	5	Marquardt et al., 2019
PRJNA479814	developmental stages	Q208 and KQ228	root/leaf/ internode	3	Thirugnanasambandam et al., 2019
PRJNA483518*	cold stress	Guitang08-1,180 and ROC22	leaf	4	Tang et al., 2018
PRJNA533093	low nitrogen	Badila	leaf/root	3	Yang et al., 2019
PRJNA291816	smut disease	RB925345	bud	3	Bedre et al., 2019
PRJNA415122	*S. scitamineum* infection	CP74_2005	bud	3	McNeil et al., 2018
PRJNA371469	osmotic stress	*S. officinarum*	root/leaf	2	Pereira-Santana et al., 2017
PRJNA681593	sucrose accumulation	Hybrids	top and bottom internodes	3	Aono et al., 2021

Represented by multiple accessions (see more details in **Supplementary Table 1).*
[1]Number of biological replicates.

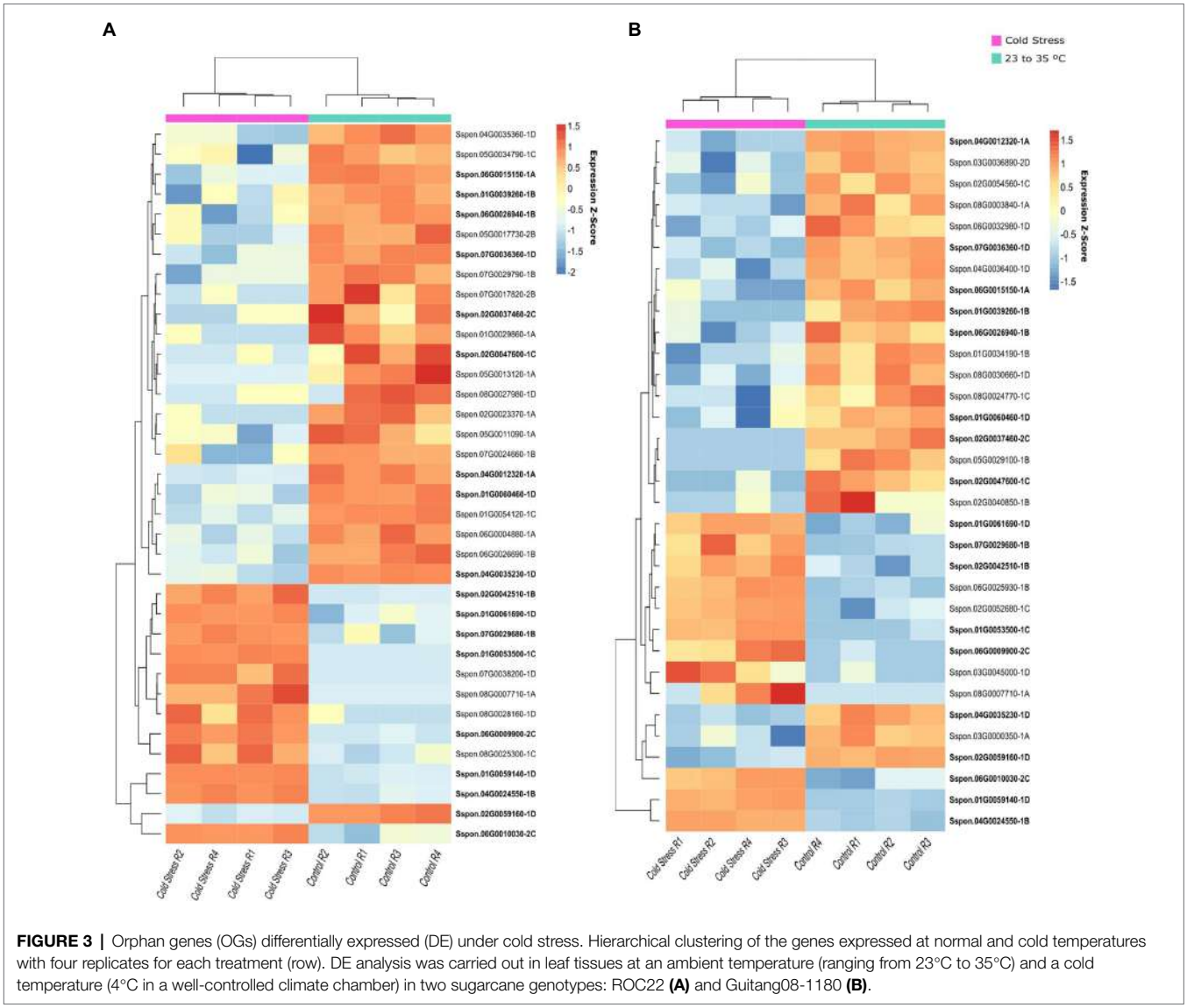

FIGURE 3 | Orphan genes (OGs) differentially expressed (DE) under cold stress. Hierarchical clustering of the genes expressed at normal and cold temperatures with four replicates for each treatment (row). DE analysis was carried out in leaf tissues at an ambient temperature (ranging from 23°C to 35°C) and a cold temperature (4°C in a well-controlled climate chamber) in two sugarcane genotypes: ROC22 **(A)** and Guitang08-1180 **(B)**.

(GO:0032446; module containing 13 OGs) and regulation of DNA methylation (GO:0044030; 10 OGs) as well as two modules with eight OGs each, which were associated with general processes such as macromolecule metabolic process (GO:0043170) and catabolic process (GO:0009056). Interestingly, a total of 24 OGs DE under stress conditions were included in modules enriched in genes functionally associated with response to stimuli, such as response to nutrient levels (GO:0031667), response to fungus (GO:0009620), and response to stress (GO:0006950).

DISCUSSION

Identification and Characterization of Sugarcane OGs

The definition and determination of OGs are context dependent, but generally, a gene that lacks detectable sequence homology in other taxa is typically classified as an OG (Rödelsperger et al., 2019). OGs with no homology to genes in other species are candidates for the *de novo* evolution of genes. However, we cannot reject the hypothesis that, in some cases, these homolog genes were missing in these target species because the genome is incomplete, missing sequences that could be biologically informative, including entire genes. Additionally, most OGs identified in our study did not show sequence similarity to any known functional domain. There are two possible reasons for this, which are linked to OG origin, complete divergence from ancestral sequences or *de novo* emergence from non-genic sequences (Van Oss and Carvunis, 2019; Singh and Syrkin Wurtele, 2020; Vakirlis et al., 2020). In the first scenario, a pair of genes sharing a common ancestor can diverge at some point when similarity is no longer detectable. Because of the lack of similarity with known functional domains, it is also possible that these genes are ncRNAs. Even though we did not observe any OGs similar to ncRNAs deposited in databases, some OGs were predicted

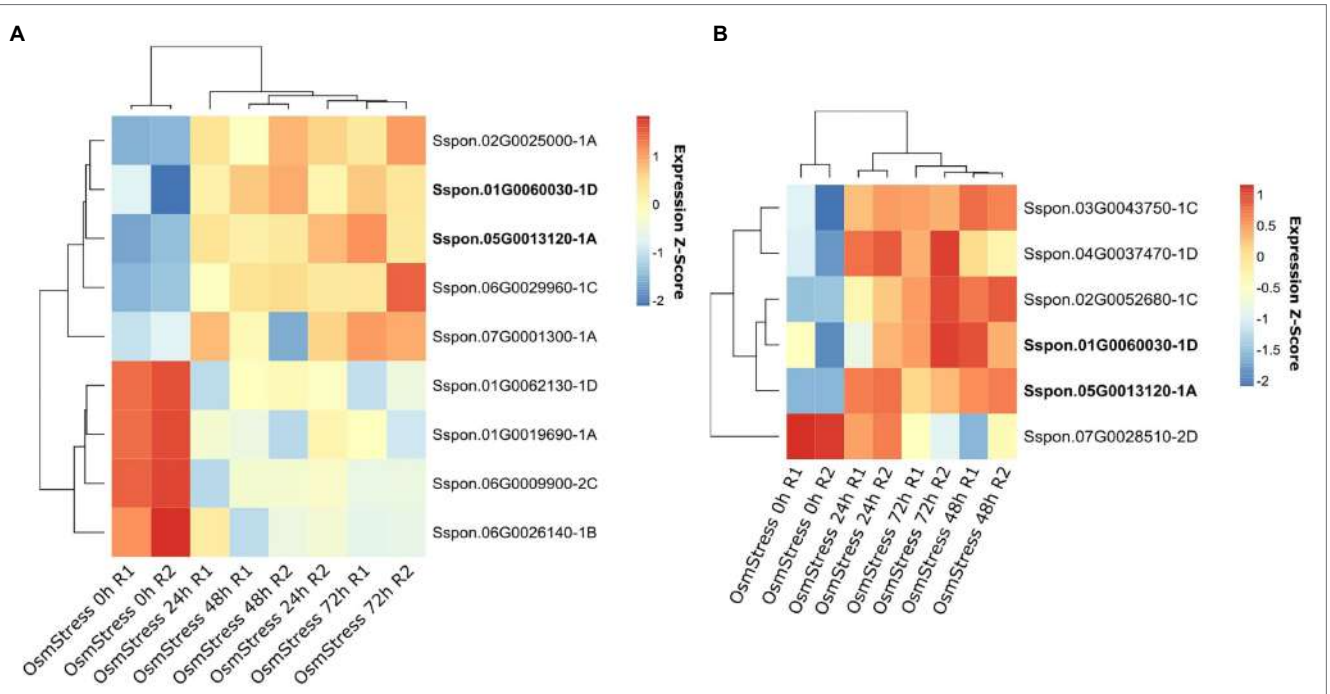

FIGURE 4 | Heatmap of orphan genes (OGs) differentially expressed (DE) under osmotic stress. Plants were subjected to osmotic stress for 24, 48, and 72 h, and other plants were maintained without osmotic stress (0 h). RNA samples were extracted from the leaves **(A)** and roots **(B)** of *S. officinarum*.

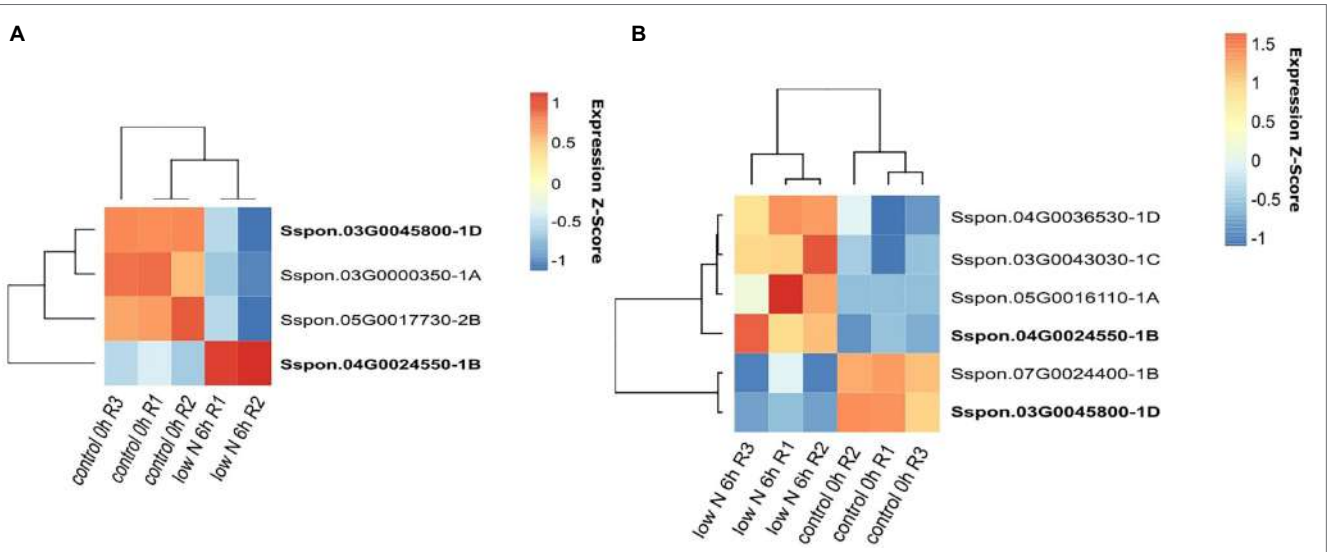

FIGURE 5 | Genes differentially expressed (DE) under low-nitrogen conditions. RNA samples were extracted from leaves of the ROC22 **(A)** and Badila **(B)** genotypes.

to be lncRNAs based on the coding potential calculator. However, this method assumes that a protein-coding RNA is more likely to have a long open reading frame than a non-coding transcript. This likely explains the strong positive correlation that we observed between gene length and the probability of being a coding sequence. Consequently, short OGs tend to be assigned as lncRNAs.

TEs Might Be Involved in OGs Origin in Sugarcane

The presence of TE fragments within a coding region is not surprising. One of the molecular mechanisms by which TEs are recruited to be part of a gene is named exonization. It takes place after TE insertion into an intron, after which parts of the TE can be incorporated, leading to the presence of a

TE exon in a protein-coding gene (Schrader and Schmitz, 2018). TE-derived proteins have been recurrently domesticated during evolution, and they have contributed to adaptive evolutionary innovation (Jangam et al., 2017; Schrader and Schmitz, 2018). In support of this view, a mechanism of OG birth *via* TEs has been reported in rice (Jin et al., 2019). Similarly, most OGs predicted in primates were found to include fragments of TEs in the transcript (Toll-Riera et al., 2008). In the sugarcane genome, we observed that most genes, both orphan and non-orphan, contained TE traces producing poor alignment in the coding region. This may indicate frequent recruitment of TEs as part of novel genes.

OGs Expression Patterns and Responses to Stress Conditions

The expression profiles of OGs across sugarcane hybrids may indicate that these genes have similar expression patterns when we compare genotypes with the same geographical origin. Indeed, it would be expected that hybrids from the same breeding programme would show greater genetic similarity, including regulatory elements, because of the admixture among genotypes sharing a common parentage. We also observed that the expression pattern of the hybrids was more similar to that of *S. officinarum* than to those of other species, suggesting preserved regulatory control of the expression of these genes after hybridization. These findings also suggest that genotype clustering is influenced by parental genome contributions. Indeed, all modern sugarcane varieties originated from hybridization between *S. officinarum* and *S. spontaneum*, followed by several backcrosses using *S. officinarum* to restore a high sucrose content (Price, 1961). This breeding process resulted in unequal contributions of the sub-genomes in *Saccharum* hybrids. For instance, the genome of the R570 hybrid received approximately 80% of its chromosomes from *S. officinarum* (D'Hont, 2005), which could explain the similar expression patterns between hybrids and *S. officinarum*. Furthermore, the observation that some OGs have tissue-specific expression may indicate a sophisticated mechanism controlling their expression, perhaps mediated by TEs. Previous studies revealed that TEs could contribute to the emergence of new regulatory elements in a tissue-specific manner (Feschotte, 2008; Sundaram et al., 2014; Trizzino et al., 2018), which is a theory based on Barbara McClintock's discovery that TEs can control gene expression (McClintock, 1956). This is a plausible conjecture because more than 70% of the sugarcane genome is represented by TEs, and fragments of these elements were found in some OGs.

Evidence of OGs being regulated under stress conditions has been reported previously. Studies have demonstrated the importance of lineage-specific genes in plants subjected to biotic and abiotic stresses. For example, in cowpea (*V. unguiculata*), OGs seem to be more involved than conserved genes in drought adaptation, as OG expression was highly induced compared to conserved gene expression under drought conditions (Li et al., 2019). Similarly, the expression levels of two OGs, CpCRP1 and CpEDR1, are modulated when individuals of a model plant widely studied for understanding the mechanism of desiccation are subjected to dehydration and rehydration processes (Giarola et al., 2014). Here, most sugarcane OGs were DE in experiments in which plants were exposed to abiotic stress. However, we cannot confirm that all these genes changed their expression pattern as an adaptive response to these stresses. Further investigation needs to be carried out to experimentally validate the effectiveness of these genes for minimizing stress effects. Even though these findings were not experimentally validated, to confirm that these genes were DE, we observed a significant number of OGs that were up- and downregulated simultaneously in independent genotypes and analyses in the same experiment.

Modules Containing OGs Are Functionally Enriched With Stress-Response Genes

The functional prediction of OGs based on sequence homology is not possible. Hence, we built a co-expression network to provide some insight into the functionality of these genes. This prediction method follows the guilty-by-association rationale (Zhang and Horvath, 2005), where the functional enrichment of modules containing genes co-expressed with OGs suggests the potential biological roles of those OGs. In general, we observed OGs distributed in several network modules, indicating that these genes are involved in diverse biological activities. This is in accordance with the findings of previous studies that described OGs as being involved in several biological networks (Beike et al., 2014; Giarola et al., 2014; Khraiwesh et al., 2015; Schlötterer, 2015). Accordingly, the main biological function attributed to these genes, the stress response, is itself a complex mechanism involving multiple biological pathways (Shulaev et al., 2008; Mantri et al., 2011). We highlight the evidence that most modules containing OGs were functionally enriched in biological processes related to stimulus responses, including those associated with various stresses. In addition to canonical terms related to stresses, such as defence responses and responses to stimuli, we also detected modules enriched, for example, with genes associated with sulphate transport and responses to auxin. Genes functionally related to these processes play an important role in stress responses (Rahman, 2012; Chan et al., 2013; Shani et al., 2017).

Because sugarcane OGs have no homology with genes in other organisms, we cannot infer their functions based on homology searches. In fact, in most cases, we did not find a known domain in the OG sequences, suggesting that sequence divergence is not the main source of OG origin in sugarcane. Even though we did not obtain strong evidence of their functionality, we cannot discard the relevance of these genes for understanding the unique biological aspects of the *Saccharum* lineage. To advance our knowledge, further investigation needs to be undertaken to better understand how these genes effectively participate in the stress response.

To date, the potential biotechnological applications of a few OGs have been tested. The QQS gene, an *Arabidopsis* gene involved in carbon and nitrogen allocation, was introduced into the soybean genome and increased starch and protein levels in the leaves (Li and Wurtele, 2014; Li et al., 2015; O'Conner et al., 2018). OGs

were also validated *via* gene editing as vital for soluble sugar metabolism in brassicas (Jiang et al., 2020). Despite the biological relevance of the OGs, it is still unknown how many of them are functional and produce stable proteins (Schlötterer, 2015; McLysaght and Hurst, 2016). Although OGs are not essential for survival, they may play an important role in responses to environmental stresses (Arendsee et al., 2014; Ma et al., 2020).

In our study, we developed an approach for the identification of OGs in sugarcane and characterization of their expression patterns. We propose that non-coding regions might provide important genetic raw material for the functional innovation, by which novel ORFs are selected and may evolve into adaptive stress response pathways. Finally, the OGs responsive to abiotic stress in sugarcane might be good candidates for further experiments to investigate their biological functions.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

CC-S, MM, and DS conducted the experiments. CC-S and AA analysed the data. CC-S wrote the manuscript. All authors discussed the data, interpreted the results, read and edited the manuscript, and approved the final version.

FUNDING

This work was supported by grants from the Fundação de Amparo à Pesquisa de Estado de São Paulo (FAPESP, 08/52197–4), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES—Computational Biology Program 88882.160095/2013–01). CC-S received a postdoctoral fellowship from FAPESP (2015/16399–5 and BEPE 2017/26781–0); AA received a PhD fellowship from FAPESP (2019/03232–6); and CS and MM received postdoctoral fellowships from FAPESP (CS 2015/24346–9 and MM 2014/11482–9). AS received a Research Fellowship from CNPq (312777/2018–3).

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for financial support and fellowships.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.923069/full#supplementary-material

REFERENCES

Aono, A. H., Pimenta, R. J. G., Garcia, A. L. B., Correr, F. H., Hosaka, G. K., Carrasco, M. M., et al. (2021). The wild sugarcane and Sorghum Kinomes: insights Into expansion, diversification, and expression patterns. *Front. Plant Sci.* 12:8623. doi: 10.3389/fpls.2021.668623

Arendsee, Z. W., Li, L., and Wurtele, E. S. (2014). Coming of age: orphan genes in plants. *Trends Genet.* 19, 698–708. doi: 10.1016/j.tplants.2014.07.003

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* 6:11. doi: 10.1186/s13100-015-0041-9

Barter, R. L., and Yu, B. (2018). Superheat: an R package for creating beautiful and extendable heatmaps for visualizing complex data. *J. Comput. Graph. Stat.* 27, 910–922. doi: 10.1080/10618600.2018.1473780

Bedre, R., Irigoyen, S., Schaker, P. D. C., Monteiro-Vitorello, C. B., Da Silva, J. A., and Mandadi, K. K. (2019). Genome-wide alternative splicing landscapes modulated by biotrophic sugarcane smut pathogen. *Sci. Rep.* 9:8876. doi: 10.1038/s41598-019-45184-1

Beike, A. K., Lang, D., Zimmer, A. D., Wüst, F., Trautmann, D., Wiedemann, G., et al. (2014). Insights from the cold transcriptome of *Physcomitrella patens*: global specialization pattern of conserved transcriptional regulators and identification of orphan genes involved in cold acclimation. *New Phytol.* 205, 869–881. doi: 10.1111/nph.13004

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Campbell, M. A., Zhu, W., Jiang, N., Lin, H., Ouyang, S., Childs, K. L., et al. (2007). Identification and characterization of lineage-specific genes within the *Poaceae*. *Plant Physiol.* 145, 1311–1322. doi: 10.1104/pp.107.104513

Cardoso-Silva, C. B., Costa, E. A., Mancini, M. C., Balsalobre, T. W. A., Canesin, L. E. C., Pinto, L. R., et al. (2014). De novo assembly and transcriptome analysis of contrasting sugarcane varieties. *PLoS One* 9:e88462. doi: 10.1371/journal.pone.0088462

Carver, T., Harris, S. R., Berriman, M., Parkhill, J., and McQuillan, J. A. (2011). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* 28, 464–469. doi: 10.1093/bioinformatics/btr703

Chan, K. X., Wirtz, M., Phua, S. Y., Estavillo, G. M., and Pogson, B. J. (2013). Balancing metabolites in drought: the sulfur assimilation conundrum. *Trends Genet.* 18, 18–29. doi: 10.1016/j.tplants.2012.07.005

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

D'Hont, A. (2005). Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. *Cytogenet. Genome Res.* 109, 27–33. doi: 10.1159/000082378

Domazet-Lošo, T., Brajković, J., and Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* 23, 533–539. doi: 10.1016/j.tig.2007.08.014

Dong, X.-M., Pu, X.-J., Zhou, S.-Z., Li, P., Luo, T., Chen, Z.-X., et al. (2022). Orphan gene PpARDT positively involved in drought tolerance potentially by enhancing ABA response in Physcomitrium (Physcomitrella) patens. *Plant Sci.* 319:111222. doi: 10.1016/j.plantsci.2022.111222

Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H., and Spillane, C. (2011). Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. *BMC Evol. Biol.* 11, 1–23. doi: 10.1186/1471-2148-11-47

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* 9, 397–405. doi: 10.1038/nrg2337

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367

Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* 8, 967–974. doi: 10.1101/gr.8.9.967

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Garsmeur, O., Droc, G., Antonise, R., Grimwood, J., Potier, B., Aitken, K., et al. (2018). A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nat. Commun.* 9:2638. doi: 10.1038/s41467-018-05051-5

Giarola, V., Krey, S., Frerichs, A., and Bartels, D. (2014). Taxonomically restricted genes of *Craterostigma plantagineum* are modulated in their expression during dehydration and rehydration. *Planta* 241, 193–208. doi: 10.1007/s00425-014-2175-2

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2011). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Guo, W.-J., Li, P., Ling, J., and Ye, S.-P. (2007). Significant comparative characteristics between orphan and nonorphan genes in the rice (Oryza sativa L.) genome. *Comp. Funct. Genom* 2007:21676. doi: 10.1155/2007/21676

Hoang, N. V., Furtado, A., Mason, P. J., Marquardt, A., Kasirajan, L., Thirugnanasambandam, P. P., et al. (2017). A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. *BMC Genomics* 18:395. doi: 10.1186/s12864-017-3757-8

Jangam, D., Feschotte, C., and Betrán, E. (2017). Transposable element domestication As an adaptation to evolutionary conflicts. *Trends Genet.* 33, 817–831. doi: 10.1016/j.tig.2017.07.011

Jiang, M., Zhan, Z., Li, H., Dong, X., Cheng, F., and Piao, Z. (2020). *Brassica rapa* orphan genes largely affect soluble sugar metabolism. *Hortic. Res.* 7:181. doi: 10.1038/s41438-020-00403-z

Jin, G. H., Zhou, Y. L., Yang, H., Hu, Y. T., Shi, Y., Li, L., et al. (2019). Genetic innovations: transposable element recruitment and de novo formation lead to the birth of orphan genes in the rice genome. *J. Syst. Evol.* 59, 341–351. doi: 10.1111/jse.12548

Jurka, J., Klonowski, P., Dagman, V., and Pelton, P. (1996). Censor—a program for identification and elimination of repetitive elements from DNA sequences. *Computers & Chemistry* 20, 119–121. doi: 10.1016/s0097-8485(96)80013-1

Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., et al. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* 45, W12–W16. doi: 10.1093/nar/gkx428

Kaur, N., Chen, W., Zheng, Y., Hasegawa, D. K., Ling, K.-S., Fei, Z., et al. (2017). Transcriptome analysis of the whitefly, *Bemisia tabaci* MEAM1 during feeding on tomato infected with the crinivirus, tomato chlorosis virus, identifies a temporal shift in gene expression and differential regulation of novel orphan genes. *BMC Genomics* 18:370. doi: 10.1186/s12864-017-3751-1

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., and Bosch, T. C. G. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet.* 25, 404–413. doi: 10.1016/j.tig.2009.07.006

Khraiwesh, B., Qudeimat, E., Thimma, M., Chaiboonchoe, A., Jijakli, K., Alzahmi, A., et al. (2015). Genome-wide expression analysis offers new insights into the origin and evolution of *Physcomitrella patens* stress response. *Sci. Rep.* 5:17434. doi: 10.1038/srep17434

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* 9:559. doi: 10.1186/1471-2105-9-559

Langfelder, P., Zhang, B., and Horvath, S. (2007). Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. *Bioinformatics* 24, 719–720. doi: 10.1093/bioinformatics/btm563

Li, G., Wu, X., Hu, Y., Muñoz-Amatriaín, M., Luo, J., Zhou, W., et al. (2019). Orphan genes are involved in drought adaptations and ecoclimatic-oriented selections in domesticated cowpea. *J. Exp. Bot.* 70, 3101–3110. doi: 10.1093/jxb/erz145

Li, L., and Wurtele, E. S. (2014). The QQS orphan gene of *Arabidopsis* modulates carbon and nitrogen allocation in soybean. *Plant Biotechnol. J.* 13, 177–187. doi: 10.1111/pbi.12238

Li, L., Zheng, W., Zhu, Y., Ye, H., Tang, B., Arendsee, Z. W., et al. (2015). QQS orphan gene regulates carbon and nitrogen partitioning across species via NF-YC interactions. *Proc. Natl. Acad. Sci. U. S. A.* 112, 14734–14739. doi: 10.1073/pnas.1514670112

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Ma, S., Yuan, Y., Tao, Y., Jia, H., and Ma, Z. (2020). Identification, characterization and expression analysis of lineage-specific genes within *Triticeae*. *Genomics* 112, 1343–1350. doi: 10.1016/j.ygeno.2019.08.003

Mantri, N., Patade, V., Penna, S., Ford, R., and Pang, E. (2011). "Abiotic stress responses in plants: present and future," in *Abiotic Stress Responses in Plants*. eds. P. Ahmad and M. Prasad (New York: Springer), 1–19.

Marquardt, A., Henry, R. J., and Botha, F. C. (2019). Midrib sucrose accumulation and sugar transporter gene expression in YCS-affected sugarcane leaves. *Trop. Plant Biol.* 12, 186–205. doi: 10.1007/s12042-019-09221-7

McClintock, B. (1956). Controlling elements and the gene. *Cold Spring Harb. Symp. Quant. Biol.* 21, 197–216. doi: 10.1101/sqb.1956.021.01.017

McLysaght, A., and Hurst, L. D. (2016). Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* 17, 567–578. doi: 10.1038/nrg.2016.78

McNeil, M. D., Bhuiyan, S. A., Berkman, P. J., Croft, B. J., and Aitken, K. S. (2018). Analysis of the resistance mechanisms in sugarcane during *Sporisorium scitamineum* infection using RNA-seq and microscopy. *PLoS One* 13:e0197840. doi: 10.1371/journal.pone.0197840

Ming, R., Liu, S.-C., Lin, Y.-R., da Silva, J., Wilson, W., Braga, D., et al. (1998). Detailed alignment of *Saccharum* and Sorghum chromosomes: comparative Organization of Closely Related Diploid and Polyploid Genomes. *Genetics* 150, 1663–1682. doi: 10.1093/genetics/150.4.1663

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2020). Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913

Mitros, T., Session, A. M., James, B. T., Wu, G. A., Belaffif, M. B., Clark, L. V., et al. (2020). Genome biology of the paleotetraploid perennial biomass crop Miscanthus. *Nat. Commun.* 11:5442. doi: 10.1038/s41467-020-18923-6

O'Conner, S., Neudorf, A., Zheng, W., Qi, M., Zhao, X., Du, C., et al. (2018). "From arabidopsis to crops: the arabidopsis QQS orphan gene modulates nitrogen allocation across species," in *Engineering Nitrogen Utilization in Crop Plants*. eds. A. Shrawat, A. Zayed and D. Lightfoot (Cham: Springer), 95–117.

Paterson, A. H., Wang, X., Li, J., and Tang, H. (2012). "Ancient and recent polyploidy in monocots," in *Polyploidy and Genome Evolution*. eds. P. S. Soltis and D. E. Soltis (Berlin: Springer), 93–108.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. doi: 10.1038/nmeth.4197

Paytuví Gallart, A., Hermoso Pulido, A., Martínez, A., de Lagrán, I., Sanseverino, W., and Aiese Cigliano, R. (2015). GREENC: a wiki-based database of plant lncRNAs. *Nucleic Acids Res.* 44, D1161–D1166. doi: 10.1093/nar/gkv1215

Pereira-Santana, A., Alvarado-Robledo, E. J., Zamora-Briseño, J. A., Ayala-Sumuano, J. T., Gonzalez-Mendoza, V. M., Espadas-Gil, F., et al. (2017). Transcriptional profiling of sugarcane leaves and roots under progressive osmotic stress reveals a regulated coordination of gene expression in a spatiotemporal manner. *PLoS One* 12:e0189271. doi: 10.1371/journal.pone.0189271

Perochon, A., Jianguang, J., Kahla, A., Arunachalam, C., Scofield, S. R., Bowden, S., et al. (2015). TaFROG encodes a Pooideae orphan protein that interacts with SnRK1 and enhances resistance to the mycotoxigenic fungus Fusarium graminearum. *Plant Physiol.* 169:2895, –2906. doi: 10.1104/pp.15.01056

Piriyapongsa, J., Kaewprommal, P., Vaiwsri, S., Anuntakarun, S., Wirojsirasak, W., Punpee, P., et al. (2018). Uncovering full-length transcript isoforms of sugarcane cultivar Khon Kaen 3 using single-molecule long-read sequencing. *PeerJ* 6:e5818. doi: 10.7717/peerj.5818

Prabh, N., and Rödelsperger, C. (2016). Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? *BMC Bioinform.* 17:226. doi: 10.1186/s12859-016-1102-x

Price, S. (1961). Cytological studies in *saccharum* and allied genera VII. Maternal chromosome transmission by *S. officinarum* in intra- and interspecific crosses. *Bot. Gaz.* 122, 298–305. doi: 10.1086/336118

Rahman, A. (2012). Auxin: a regulator of cold stress response. *Physiol. Plant.* 147, 28–35. doi: 10.1111/j.1399-3054.2012.01617.x

Renny-Byfield, S., and Wendel, J. F. (2014). Doubling down on genomes: polyploidy and crop plants. *Am. J. Bot.* 101, 1711–1725. doi: 10.3732/ajb.1400119

Rödelsperger, C., Prabh, N., and Sommer, R. J. (2019). New gene origin and deep taxon phylogenomics: opportunities and challenges. *Trends Genet.* 35, 914–922. doi: 10.1016/j.tig.2019.08.007

Schlötterer, C. (2015). Genes from scratch--the evolutionary fate of de novo genes. *Trends Genet.* 31, 215–219. doi: 10.1016/j.tig.2015.02.007

Schrader, L., and Schmitz, J. (2018). The impact of transposable elements in adaptive evolution. *Mol. Ecol.* 28, 1537–1549. doi: 10.1111/mec.14794

Shani, E., Salehin, M., Zhang, Y., Sanchez, S. E., Doherty, C., Wang, R., et al. (2017). Plant stress tolerance requires auxin-sensitive Aux/IAA transcriptional repressors. *Curr. Biol.* 27, 437–444. doi: 10.1016/j.cub.2016.12.016

Shulaev, V., Cortes, D., Miller, G., and Mittler, R. (2008). Metabolomics for plant stress response. *Physiol. Plant.* 132, 199–208. doi: 10.1111/j.1399-3054.2007.01025.x

Singh, U., and Syrkin Wurtele, E. (2020). How new genes are born. *elife* 9:e55136. doi: 10.7554/eLife.55136

Soltis, D. E., Albert, V. A., Leebens-Mack, J., Bell, C. D., Paterson, A. H., Zheng, C., et al. (2009). Polyploidy and angiosperm diversification. *Am. J. Bot.* 96, 336–348. doi: 10.3732/ajb.0800079

Souza, G. M., Van Sluys, M.-A., Lembke, C. G., Lee, H., Margarido, G. R. A., Hotta, C. T., et al. (2019). Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. *GigaScience* 8:giz129. doi: 10.1093/gigascience/giz129

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., et al. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* 24, 1963–1976. doi: 10.1101/gr.168872.113

Szcześniak, M. W., Rosikiewicz, W., and Makałowska, I. (2015). CANTATAdb: A collection of plant long non-coding RNAs. *Plant Cell Physiol.* 57:e8. doi: 10.1093/pcp/pcv201

Tang, S., Yang, L., and Li, Y. (2018). Comparative analysis on Transcriptome Among different sugarcane cultivars Under low temperature stress[J]. *Biotechnol. Bull.* 34, 116–124. doi: 10.13560/j.cnki.biotech.bull.1985.2018-0522

Tautz, D., and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12, 692–702. doi: 10.1038/nrg3053

Thirugnanasambandam, P. P., Hoang, N. V., and Henry, R. J. (2018). The challenge of analyzing the sugarcane genome. *Front. Plant Sci.* 9:616. doi: 10.3389/fpls.2018.00616

Thirugnanasambandam, P. P., Mason, P. J., Hoang, N. V., Furtado, A., Botha, F. C., and Henry, R. J. (2019). Analysis of the diversity and tissue specificity of sucrose synthase genes in the long read transcriptome of sugarcane. *BMC Plant Biol.* 19:160. doi: 10.1186/s12870-019-1733-y

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., et al. (2008). Origin of primate orphan genes: A comparative genomics approach. *Mol. Biol. Evol.* 26, 603–612. doi: 10.1093/molbev/msn281

Trizzino, M., Kapusta, A., and Brown, C. D. (2018). Transposable elements generate regulatory novelty in a tissue-specific fashion. *BMC Genomics* 19:468. doi: 10.1186/s12864-018-4850-3

Vakirlis, N., Carvunis, A.-R., and McLysaght, A. (2020). Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *elife* 9:e53500. doi: 10.7554/eLife.53500

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* 10, 725–732. doi: 10.1038/nrg2600

Van Oss, S. B., and Carvunis, A.-R. (2019). De novo gene birth. *PLoS Genet.* 15:e1008160. doi: 10.1371/journal.pgen.1008160

Wheeler, T. J., and Eddy, S. R. (2013). Nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29, 2487–2489. doi: 10.1093/bioinformatics/btt403

Wilson, G. A., Bertrand, N., Patel, Y., Hughes, J. B., Feil, E. J., and Field, D. (2005). Orphans as taxonomically restricted and ecologically important genes. *Microbiology* 151, 2499–2501. doi: 10.1099/mic.0.28146-0

Xu, Y., Wu, G., Hao, B., Chen, L., Deng, X., and Xu, Q. (2015). Identification, characterization and expression analysis of lineage-specific genes within sweet orange (*Citrus sinensis*). *BMC Genomics* 16:995. doi: 10.1186/s12864-015-2211-z

Yang, Y., Gao, S., Su, Y., Lin, Z., Guo, J., Li, M., et al. (2019). Transcripts and low nitrogen tolerance: regulatory and metabolic pathways in sugarcane under low nitrogen stress. *Environ. Exp. Bot.* 163, 97–111. doi: 10.1016/j.envexpbot.2019.04.010

Yip, A. M., and Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinform.* 8:22. doi: 10.1186/1471-2105-8-22

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:1128. doi: 10.2202/1544-6115.1128

Zhang, X., Xuan, J., Yao, C., Gao, Q., Wang, L., Jin, X., et al. (2022). A deep learning approach for orphan gene identification in moso bamboo (*Phyllostachys edulis*) based on the CNN+transformer model. *BMC Bioinfo.* 23:162. doi: 10.1186/s12859-022-04702-1

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2018). Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* 50, 1565–1573. doi: 10.1038/s41588-018-0237-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.