Vigna et al. BMC Genomics (2016) 17:910
DOI 10.1186/s12864-016-3270-5

BMC Genomics

RESEARCH ARTICLE Open Access

# Leaf transcriptome of two highly divergent genotypes of *Urochloa humidicola* (Poaceae), a tropical polyploid forage grass adapted to acidic soils and temporary flooding areas

Bianca Baccili Zanotto Vigna[1†], Fernanda Ancelmo de Oliveira[2†], Guilherme de Toledo-Silva[2,5†], Carla Cristina da Silva[2], Cacilda Borges do Valle[3] and Anete Pereira de Souza[2,4*]

## Abstract

**Background:** *Urochloa humidicola* (Koronivia grass) is a polyploid (6x to 9x) species that is used as forage in the tropics. Facultative apospory apomixis is present in most of the genotypes of this species, although one individual has been described as sexual. Molecular studies have been restricted to molecular marker approaches for genetic diversity estimations and linkage map construction. The objectives of the present study were to describe and compare the leaf transcriptome of two important genotypes that are highly divergent in terms of their phenotypes and reproduction modes: the sexual BH031 and the aposporous apomictic cultivar BRS Tupi.

**Results:** We sequenced the leaf transcriptome of Koronivia grass using an Illumina GAIIx system, which produced 13.09 Gb of data that consisted of 163,575,526 paired-end reads between the two libraries. We *de novo*-assembled 76,196 transcripts with an average length of 1,152 bp and filtered 35,093 non-redundant unigenes. A similarity search against the non-redundant National Center of Biotechnology Information (NCBI) protein database returned 65 % hits. We annotated 24,133 unigenes in the Phytozome database and 14,082 unigenes in the UniProtKB/Swiss-Prot database, assigned 108,334 gene ontology terms to 17,255 unigenes and identified 5,324 unigenes in 327 known metabolic pathways. Comparisons with other grasses via a reciprocal BLAST search revealed a larger number of orthologous genes for the *Panicum* species. The unigenes were involved in C4 photosynthesis, lignocellulose biosynthesis and flooding stress responses. A search for functional molecular markers revealed 4,489 microsatellites and 560,298 single nucleotide polymorphisms (SNPs). A quantitative real-time PCR analysis validated the RNA-seq expression analysis and allowed for the identification of transcriptomic differences between the two evaluated genotypes. Moreover, 192 unannotated sequences were classified as containing complete open reading frames, suggesting that the new, potentially exclusive genes should be further investigated.

(Continued on next page)

* Correspondence: anete@unicamp.br
†Equal contributors
[2]Center for Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Campinas, SP, Brazil
[4]Department of Plant Biology, Biology Institute, UNICAMP, Campinas, SP, Brazil
Full list of author information is available at the end of the article

(Continued from previous page)

**Conclusion:** The present study represents the first whole-transcriptome sequencing of *U. humidicola* leaves, providing an important public information source of transcripts and functional molecular markers. The qPCR analysis indicated that the expression of certain transcripts confirmed the differential expression observed *in silico*, which demonstrated that RNA-seq is useful for identifying differentially expressed and unique genes. These results corroborate the findings from previous studies and suggest a hybrid origin for BH031.

**Keywords:** *Brachiaria*, *de novo* transcriptome assembly, molecular markers, RNA-seq, tropical grasses

## Background

Tropical pastures comprise plants from several genera of grasses and legumes [1]. Brazil has the second largest effective cattle herd and the highest beef exports worldwide [2], with 190 million ha of pastures [3]. However, most of the areas cultivated with forage crops are established with a limited number of exotic and clonal reproduction cultivars [4], representing a high risk of genetic vulnerability for forage production.

To minimize the risk of planting large contiguous areas and diversify Brazilian pastures, genetically improved forage species and new cultivars must be established [5]. The *Urochloa* (Poaceae) genus is primarily native to tropical African savannas and widely used for pastures in the tropics. This genus comprises approximately 120 identified species [6], with the species *U. brizantha*, *U. decumbens*, *U. ruziziensis* and *U. humidicola* accounting for 85 % of the cultivated pastures in Brazil [1].

*Urochloa humidicola* (Rendle) Morrone & Zuloaga (syn. *Brachiaria humidicola* (Rendle) Schweick) [7] is widely used in the pastures of Brazil and elsewhere in the tropics, particularly in acidic and poorly drained areas [8]. Although the interest of producers in this species has increased, few cultivars, such as Tully, Llanero and BRS Tupi, are available in Brazil. The development and adoption of new Koronivia grass cultivars with a broad genetic base is crucial for the diversification of forage pastures in the tropics and the enhancement of forage production. Thus, hybrids are being selected for release in this country [9, 10], and they are all derived from the sexual genotype BH031 and the facultative apomictic cultivar BRS Tupi, which are two highly divergent genotypes of Koronivia grass [11, 12].

Koronivia grass is an outcrossed and wind-pollinated perennial tropical grass that primarily reproduces through facultative apomixis, although a single sexual genotype has been identified [12, 13]. This species shows variable levels of ploidy (6x to 9x) and a basic chromosome number of x = 6 [14, 15]. This species presents a relatively large and complex genome (1953 Mbp) [16], although limited data are available on the function and structure of this genetic material.

The genetic and genomic influences on the agronomic traits of interest and the mechanisms involved in the genotype-phenotype relationships in this species are poorly understood [17]. A search for *Urochloa* in the National Center of Biotechnology Information (NCBI) database revealed 2,237 expressed sequence tags (ESTs) as of March 2015. The generation of a reference transcriptome is a critical step for exploring the molecular machinery of a species with few genomic resources, such as *U. humidicola*. RNA-seq is an efficient method of analyzing transcriptomes, generating comprehensive and in-depth biological resources [18, 19] and providing new insights into gene expression patterns. High-throughput sequencing methods produce millions of short sequence reads that facilitate the profiling of gene expression, the discovery of novel transcribed regions, the detection of alternative splicing isoforms, and the identification of valuable molecular markers, such as microsatellites and single nucleotide polymorphisms (SNPs). Therefore, these methods have the potential for use in molecular breeding strategies to improve the productivity and quality of forage [20].

In the present study, we describe the initial sequencing and *de novo* assembly of the transcriptome of *U. humidicola* leaves from two highly divergent genotypes (the sexual accession BH031 and the facultative apomictic cultivar BRS Tupi) that are important for the species biology and breeding using an Illumina massive parallel sequencing platform. The objectives of the present study were to provide novel genetic resources for the species through the *de novo* assembly of RNA-seq data from the leaves of two divergent genotypes, assess the transcriptomic differences between the genotypes and identify potential functional molecular markers because the genomic single sequence repeat (SSR) developed for *U. humidicola* [11, 21–23] displays amplification issues in the sexual genotype BH031.

## Methods

### Plant materials and DNA and RNA extraction

Two divergent genotypes of *U. humidicola* were sampled for RNA sequencing: the sexual accession BH031 and the facultative apomictic cultivar BRS Tupi. BRS Tupi stands out as an option for low-fertility pasture soils subjected to temporary flooding [24]. In addition to their modes of reproduction, both genotypes differ in terms of

growth habit, tillering intensity, leaf width and productivity as shown in Fig. 1. These genotypes are the genitors of a mapping population at Embrapa Beef Cattle in Campo Grande, MS, where the genotypes were sampled under previously described conditions [10] and at the same time of day.

On May 12, 2012 (rainy season in Brazil), young leaves were collected from each genotype and immediately placed in RNAlater (Life Technologies, Carlsbad, CA, USA) for stabilization and storage. The samples were transported to the Multiuser Laboratory of Genotyping and Sequencing (LMGS) at the Center of Molecular Biology and Genetic Engineering (CBMEG/UNICAMP, SP, Brazil) and stored at −80 °C until further use. Total RNA was isolated using a modified lithium chloride protocol [25]. We assessed the integrity and quantity of the obtained RNA using a 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA). Equal molar quantities of high-quality RNA from each material were used as templates for the cDNA synthesis in the subsequent steps.

For the quantitative RT-PCR, young leaves were collected from three different plants of each genotype on January 26, 2015. Total RNA [25] and total DNA [26] were extracted as previously described. The purity and concentration of the RNA were determined using a NanoVue Plus spectrophotometer (GE Healthcare, Piscataway, NJ, USA), and the purity and concentration of the DNA were determined using a NanoDrop 2000/2000c spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA).

### Preparation and sequencing (RNA-seq) of the cDNA library

A cDNA library was constructed separately for each sample using a TruSeq RNA Sample Preparation Kit (Illumina Inc., San Diego, CA, USA) according to the manufacturer's instructions. The libraries´ quality was confirmed using a 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA) and quantified via qPCR (Illumina protocol SY-930-10-10). Clustering was conducted using a TruSeq PE Cluster Kit on cBot (Illumina Inc., San Diego, CA, USA). Subsequently, the cDNA libraries were sequenced, with each one in a single lane, using an Illumina Genome Analyzer IIx with TruSeq SBS 36-Cycle kits (Illumina, San Diego, CA, USA) according to the manufacturer's specifications for 80-bp PE reads.

### Data filtering and *de novo* assembly

The raw data that were generated from the Illumina sequencer were initially obtained in BCL format and subsequently converted to qSeq format using the Off-Line Basecaller v.1.9.4 (OLB) software. We further converted the qSeq files into FastQ files containing 80-bp reads using a custom script. Filtering for high-quality (HQ) reads was performed using the NGS QC Toolkit 2.3 [27]. Initially, low-quality reads (Phred quality score < 20 in 75 % of all reads) and reads with fewer than 60 bases were removed. Subsequently, the remaining reads were trimmed at the 3' end, and bases with Phred scores < 20 were removed. All reads were deposited in the NCBI Short Read Archive (SRA) under accession number SRP065020, at the following link: https://www.ncbi.nlm.nih.gov/sra/?term=SRP065020.

In the present study, we compared the performance of two commonly used *de novo* assemblers. Only HQ paired-end (PE) reads from the two sequenced samples were used for assembly and combined together to obtain a representative dataset for the *U. humidicola* leaf transcriptome. We used Trinity (Trinity-r2012-10-05 version) [28] and

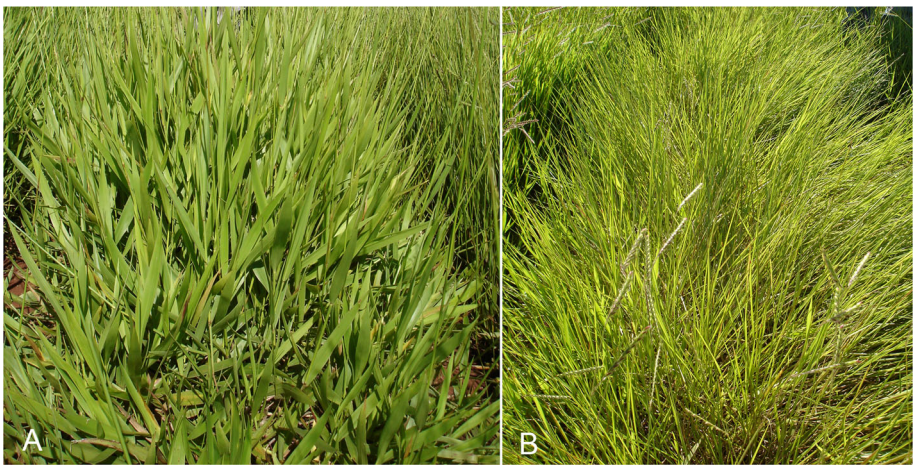


**Fig. 1** *Urochloa humidicola* genotypes sequenced. **a**) Sexual accession BH031 and **b**) apomictic cv. BRS Tupi; both are maintained at the germplasm bank at Embrapa Gado de Corte

CLC Genomics Workbench (v4.9; CLC Bio, Cambridge, MA, USA) software for the *de novo* transcriptome assembly from short reads using de Bruijn graphs, for which the reads were initially broken into smaller fragments referred to as k-mers (k denotes the length of these sequences) [28, 29] and subsequently assembled into larger sequences without using a reference genome or transcriptome. Only the default k-mer length (25-mer) was selected for the Trinity assembler, whereas two different k-mer lengths (25-mer and 45-mer) were selected for the CLC Genomics Workbench assembler. The default settings were used for both software programs. To assess the integrity of the assembled transcriptomes, HQ short reads were mapped back into the transcript dataset using the Bowtie sequence aligner [30] with default parameters. We considered various parameters, including the total number of contigs (>300 bp), percentage of mapped reads, length of N50 contigs and average length of the contigs, to select the optimal assembly.

#### Functional annotation

After a comparison between the assemblies and selection of the Trinity assembly, which was the optimal method, we further selected distinct sequences (unigenes) among the complete dataset of transcripts. The Trinity assembly has three modules that are executed in sequence (Inchworm, Chrysalis, and Butterfly), and the unigenes were filtered using criteria that retain the first Butterfly transcript generated per Chrysalis component, which is considered representative.

The unigenes were compared against the NCBI non-redundant protein (Nr) database and the UniProtKB/Swiss-Prot database [31]. A homology search against these databanks was performed using the BLASTX option from the BLAST+ suite [32] with an e-value cutoff of 1e-06.

We constructed a grass transcriptome databank that contained several species (*Brachypodium distachyon, Oryza sativa, Panicum hallii, Panicum virgatum, Sorghum bicolor,* and *Setaria italica*) using Phytozome v9.0 [33] and the previously developed *Panicum maximum* transcriptome [34]. The unigenes were compared with this grass databank using BLASTN (e-value cutoff of 1e-10) from the BLAST+ suite and a custom script to generate reciprocal BLAST hits (RBH), and the putative orthologous relationships were determined. This approach was also used to individually compare each species.

Gene ontology (GO) [35] terms were retrieved from the sequence that was functionally annotated in the NCBI nr databank using Blast2GO software [36]. The GO terms were mapped to each annotated transcript based on the 10 best hits, after which we proceeded to the annotation step in the Blast2GO software using default parameters with the exception of the e-value cutoff (1e-10). The GO annotation was enriched using ANNEX [36], and we used Go-slim [35] with plant slim (*Arabidopsis thaliana*) as an alias to summarize the GO terms of the transcriptome and facilitate data interpretation. WEGO [37] was used for the graphical representation of the GO functional classification, demonstrating the distribution of *U. humidicola* gene functions into the three main ontology categories: biological processes, molecular functions, and cellular components. An enrichment analysis using Fisher's exact test (FDR < 0.05) in the Blast2GO software was applied to search for enriched GO terms between the unique transcripts of BH031 and BRS Tupi and between the unique transcripts of *U. humidicola* (compared with the grass databank) and the assembled transcriptome in the present study. Unigenes were assigned to known metabolic pathways using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [38]. The KEGG Automatic Annotation Server (KAAS) [39] was used to obtain assignments with the bidirectional best hit (BBH) method, which generated KEGG Orthology (KO) terms related to the assembled leaf transcriptome. The sequences were matched against the Pfam database [40] using the InterProScan tool [41] to identify the protein domains of the assembled unigenes. Moreover, we searched the assembled dataset of transcripts for open reading frames (ORFs) using TransDecoder (http://transdecoder.github.io/) with a minimum length of 100 amino acids (aa) and default parameters.

#### Read mapping and transcript abundance analyses

We used RNA-seq by Expectation Maximization (RSEM) [42] to estimate the transcript expression levels. The analyses were performed using combined data from the samples and the data from each genotype separately based on the number of fragments that were mapped to the Trinity assembly contigs. The transcript FPKM values (fragments per kilobase of transcript per million mapped reads) were estimated using RSEM with Bowtie read mapping [30]. We used the RSEM analysis as the main criteria to differentiate the unigene distribution along the two sequenced genotypes, and only transcripts with FPKM values larger than 0.5 were considered to identify unique and shared transcripts. A Venn diagram was obtained using Venny 2.0.2 [43].

#### Putative molecular markers

The MISA (MIcroSAtellite) script [44] was used to identify SSRs. Microsatellite regions were defined as containing at least six motif repetitions for dinucleotides and four motif repetitions for each tri/tetra/penta/hexa-nucleotide motif. A SSR motif was defined as a compound when two or more SSR motifs were interrupted by up to 50 bp.

To identify putative SNP positions, the CLC Genomics Workbench was used to map reads to the *de novo* assembled leaf transcriptome. The parameters were as follows: length fraction = 0.9, similarity = 0.9, mismatch cost = 2, insertion cost = 3, and deletion cost = 3. Subsequently, the putative SNPs were detected using the following criteria: window length = 11, minimum coverage = 20, minimum central base quality score = 30, minimum average quality score = 20, minimum coverage at SNP site = 20, minimum frequency = 10 %, and ploidy = 6. Default settings were used for the remaining parameters.

### Quantitative RT-PCR analysis

To validate the RNA-seq expression analysis, transcripts that displayed different expression patterns *in silico* were selected for quantitative RT-PCR analysis based on the RSEM quantification. Primer3Plus software [45] was used to design primer pairs for the sequences. The target amplicon size was set at 70–150 bp, and the optimal annealing temperature was 60 °C and optimal primer length was 20 bp. Reference genes were selected according to previous studies in grasses [46–49], searched in the *U. humidicola* transcriptome, and used to design the primers as described above. All of the primer pairs were initially evaluated by PCR using genomic DNA as a template. The primer pairs that successfully amplified the DNA of both genotypes with an amplification efficiency of 90-110 % by qPCR were used for the expression analysis.

Three biological replicates for each of the two genotypes were used for the qPCR assays. Total RNA (500 ng) was used for cDNA synthesis using the QuantiTect Reverse Transcription Kit (Qiagen Inc., Chatsworth, CA, USA), which includes a genomic DNA removal step. The cDNAs were diluted (1:20) in nuclease-free water, and 2 μL samples were used for the qPCR reactions.

Quantitative RT-PCR was performed using a CFX384 Real-Time PCR Detection System with iTaq Universal SYBR® Green Supermix (Bio-Rad Laboratories Inc., Hercules, CA, USA) according to the manufacturer's instructions, and the final primer concentration was 0.3 μM. The reaction conditions were 95 °C for 10 min and then 40 cycles of 95 °C for 30 s and 60 °C for 1 min. No template controls for either primer pair were included, and each reaction was performed in triplicate. The BRS Tupi genotype sample was used as the control for the normalization of gene expression. The presence of single amplicons in the PCR products was confirmed by a melting curve analysis with temperatures ranging from 65 °C to 95 °C at increments of 0.5 °C. The baseline and Cq values were automatically determined, and the expression analysis (ΔΔCt method) was performed using CFX Manager 2.1 software (Bio-Rad Laboratories, Inc., USA).

Reference genes were selected according to the amplification efficiency (E = 90-110 % and R^2 > 0.99), gene expression stability values (M < 0.5) and coefficients of variance (CV < 0.25) among the samples. Statistical significance was tested using a pair-wise fixed reallocation randomization test (10,000 iterations) with the relative expression software tool (REST 2009, Qiagen) [50]. Differences were considered statistically significant at $p < 0.05$.

## Results and discussion

### Sequencing and filtering

In the present study, we performed a *de novo* assembly of the *U. humidicola* leaf transcriptome and characterized the transcriptome at the single nucleotide level. Each genotype was sequenced independently using the Illumina GAIIx platform (Table 1), which resulted in a total of 163,575,526 (13.09 Gb) PE reads from both libraries combined. Sequencing of the cv. BRS Tupi genotype produced a slightly larger number of reads (54.85 % of total reads) compared with that of the BH031 genotype (45.15 % of total reads). However, this difference did not interfere with the assembly or characterization of the *U. humidicola* leaf transcriptome. The short read filtering process resulted in a total of 133,201,738 (10.66 Gb) high quality (HQ) PE reads from both libraries, and these sequences were used for the *de novo* assembly. The genotype-filtered data are shown in Table 1, and they demonstrate that BRS Tupi retained more HQ reads than did BH031 after the cleaning and filtering procedures; however, this difference was not significant.

### *De novo* transcriptome assembly

We compared the results of the Trinity and CLC Genomics Workbench assemblers because these methodologies use different approaches for the *de novo* assembly of short reads. For the CLC Genomics Workbench assemblies, we observed that the total number of assembled contigs, the mean length and the N50 length displayed better metrics when using a k-mer value of k = 25 compared with k = 45 (Table 2); however, an increase in the k-mer value to k = 45 led to a slightly greater improvement in the percentage of mapped reads. The Trinity assembly displayed the largest values for the total assembled transcripts, mean length and N50 length (Table 2).

**Table 1** Summary of Illumina sequencing output statistics

| | | BH031 | BRS Tupi | Total |
|---|---|---|---|---|
| Sequenced | Number of reads | 73779964 | 89795562 | 163575526 |
| | Total bases (Gb) | 5.90 | 7.18 | 13.09 |
| Filtered | Number of reads | 57322934 | 75878804 | 133201738 |
| | Total bases (Gb) | 4.59 | 6.07 | 10.66 |
| | High quality data | 77.8 % | 84.5 % | 81.4 % |

**Table 2** Comparison of the *de novo* assemblies generated using the Trinity and CLC Genomics Workbench programs

| | Trinity | CLC Genomics Workbench | |
|---|---|---|---|
| K-mer (bp) | 25 | 25 | 45 |
| Total number of contigs | 76196 | 51291 | 44899 |
| Min contig length (bp) | 301 | 300 | 300 |
| Max contig length (bp) | 7392 | 6174 | 5506 |
| Mean contig length (bp) | 877 | 617 | 551 |
| Median contig length (bp) | 637 | 454 | 435 |
| N50 (bp) | 1152 | 665 | 562 |
| Mapped reads (%) | 82.94 | 79.53 | 80.23 |

The Bowtie aligner, which only considered properly mapped paired ends, mapped 82.94 % of the reads onto assembled sequences. Moreover, the Trinity assembler provided isoform data on the transcripts, which will be valuable for future studies. Thus, the Trinity assembly was preferable to the CLC Genomics workbench assembly.

Considering the assembly metrics that were obtained using the short reads generated using the Illumina platform, the N50 contig length of the *U. humidicola* leaf transcriptome was greater than that of *Panicum maximum* (981 bp) [34] and bamboo (1,132 bp) [51]. In addition, the average length of the unigenes was greater than that of *P. maximum* (758 bp) [34], *Salvia splendens* (772 bp) [52], *Youngia japonica* (795 bp) [53] and *Houttuynia cordata* (679 bp) [54], suggesting that the quality of the assembly from Illumina PE sequencing for this non-model organism was satisfactory compared with that of the other *de novo* assemblies.

A total of 35,093 sequences (46.12 % of all transcripts) was filtered as unigenes from the Trinity assembly using established criteria, which resulted in a dataset with a mean length of 870 bp, a N50 contig length of 1,171 bp and a GC content of 47.08 %. We obtained 7,588 (21.62 %) unigenes that were longer than 1 kb, a size range that commonly confers a high annotation rate [34, 55]. We also compared the unigene metrics obtained here to the transcriptomes that were obtained from Phytozome (Additional file 1) and found that the observed total contig number, N50 contig length, average contig length and GC content more closely resembled the Phytozome transcriptomes of these species. It is important to note that the Phytozome datasets are more curated and constantly updated as part of an ongoing process, whereas the present study is an initial attempt to characterize the *U. humidicola* leaf transcriptome.

## Annotation

We performed a unigene homology search using several databases, including the NCBI non-redundant protein database, the UniProtKB/Swiss-Prot database, a custom grass database, and the Kyoto Encyclopedia of Genes and Genomes (KEGG), and we retrieved GO terms and Pfam protein domains. Table 3 shows the overall results of the annotation. Among the 35,093 unigenes, 23,071 (65.7 %) displayed homology to sequences in the NCBI nr database. Most of these unigenes were related to proteins from *Sorghum bicolor* (9,866), followed by proteins from *Zea mays* (7,134), *Oryza sativa* (2,631), *Brachypodium distachyon* (701) and *Aegilops tauschii* (563) (Additional file 2). These five species accounted for ~90 % of the total top hits in the NCBI nr databank comparison, and this result was expected because these species are closely related to *U. humidicola* and classified within Poaceae, which includes species with a high level of genomic synteny [56]. As expected, unigenes with a size > 1 kb showed a higher annotation rate (96.6 %) compared with unigenes with a size less than 1 kb (57.2 %) (Fig. 2). Based on the NCBI nr annotation, 17,255 (49.2 %) unigenes had assigned GO terms, and these terms were divided into three main classes of ontologies: cellular component, molecular function and biological process. The majority of the GO terms were assigned to the terms cellular component (45,259 terms, 41.78 %) and biological process (41,785 terms, 38.57 %), whereas a smaller number were assigned to the term molecular function (19.64 %). Regarding the cellular component class, 13 different terms were obtained, and proteins located in the cell and cell parts were dominant and presented same number of unigenes (13,135) because these GO terms are typically assigned together. In the molecular function class, proteins related to binding (9,483) and catalytic activity (8,805) were primarily observed for the *U. humidicola* unigenes, and these classes included 14 different terms. Moreover, the biological process ontology primarily consisted of proteins involved in cellular and physiological processes, and the terms cellular process (9,722 unigenes) and metabolic process (9,707 unigenes) were the most representative GO terms followed by response to stimulus (3,470 unigenes), for which the response to stress subterm (2,216 unigenes) was the most representative category (Fig. 3).

We also identified 3,024 different Pfam protein domains in 10,880 unigenes using InterProScan. The top 10 Pfam domains are presented in Fig. 4. The most abundant protein domain was the pentatricopeptide repeat (PPR),

**Table 3** Summary of the annotated *U. humidicola* transcriptome

| | No. of annotated unigenes | % of annotated unigenes |
|---|---|---|
| NCBI nr | 23,071 | 65.7 % |
| UniProtKB/Swiss-Prot | 14,082 | 40.1 % |
| Gene Ontology | 17,255 | 49.2 % |
| KEGG | 5,324 | 15.2 % |
| Pfam | 10,880 | 31.0 % |

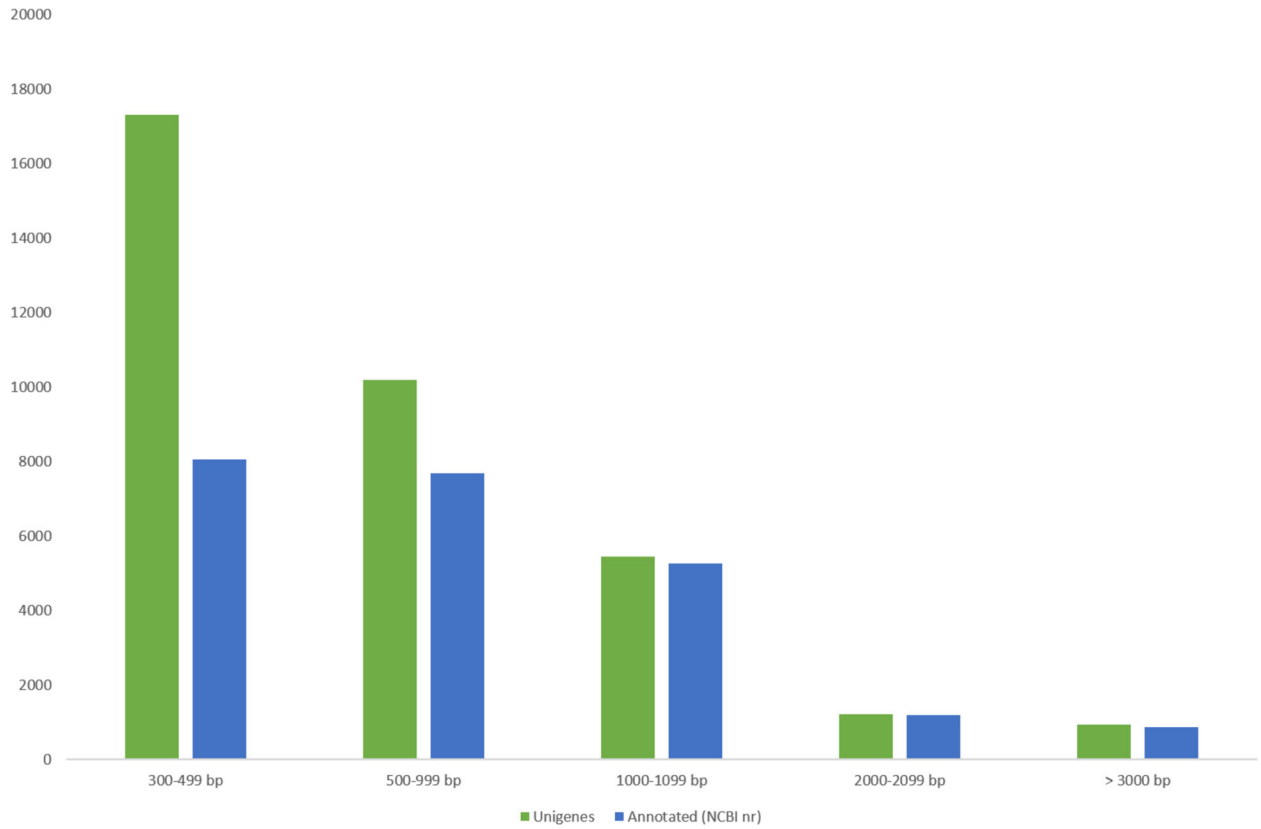


**Fig. 2** Comparison of the size ranges of the unigenes and annotated (NCBI nr) unigenes

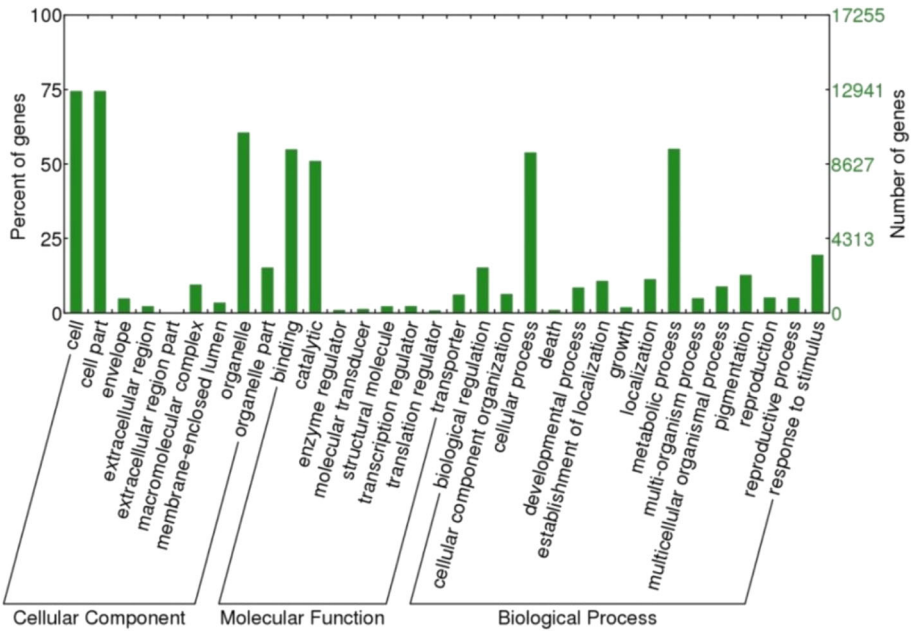


**Fig. 3** GO classification for the *U. humidicola* leaf transcriptome

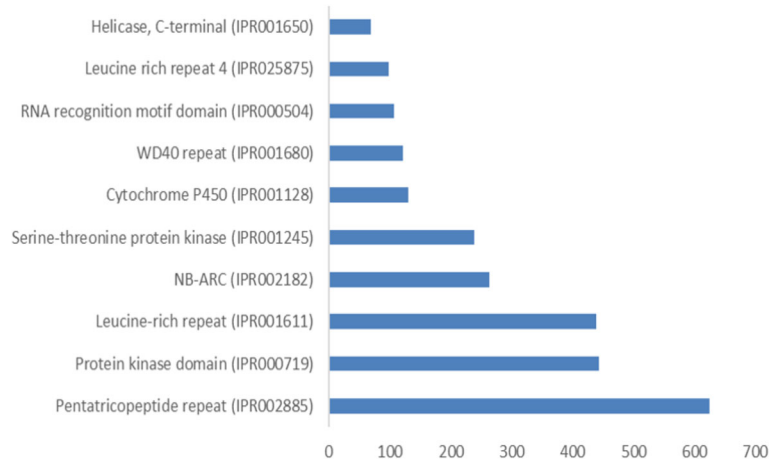


**Fig. 4** Distribution of the top 10 Pfam domains identified in the *U. humidicola* unigenes

which is characterized by tandem repeats of a degenerate 35-aa-long motif [57]. PPRs play a role in post-transcriptional processes within organelles and possess sequence-specific RNA-binding properties. Plant genomes contain between one and five hundred PPR genes per genome, whereas non-plant genomes encode two to six PPR proteins [58]. We identified 624 unigenes with PPR domains in the present study. An interesting abundant Pfam domain was NB-ARC, a signaling motif that is shared by plant resistance gene products and regulators of cell death in animals [59]. Additionally, several Cytochrome P450 (CYP450) domains were identified among the unigenes. Cytochrome P450 enzymes are a superfamily of monooxygenases that are observed in all kingdoms of life, and they play several roles in metabolism and reflect increased biochemical diversity. In plants, CYP450s are involved in the biosynthesis of several compounds, such as hormones, defensive compounds and fatty acids [60].

To correlate *U. humidicola* unigenes with known metabolic pathways, we used the KAAS server to assign sequences with KO terms and their respective KEGG maps. A total of 5,234 (15.2 %) assembled unigenes were associated with 3,936 KO terms and 327 pathways. The primarily represented pathways were metabolic pathways (750 members) and secondary metabolite biosynthesis pathways (313 members). Other important pathways included the glycolysis/gluconeogenesis pathway (Additional file 3) and plant hormone signal transduction pathway (Additional file 4). The KAAS results revealed that the assembled unigenes were distributed among several metabolic pathways.

The transcripts and unigenes were submitted to an ORF predictor using TransDecoder, and we detected ORFs in 53.85 % of the assembled transcripts and in 61.64 % of the unigenes. Further information regarding the ORFs is provided in Additional file 5.

### Transcript abundance analysis

Next-generation sequencing provided an overview of the genomic expression profile in the leaves of *U. humidicola* at a specific moment. The mean FPKM value for all of the unigenes was ~29. We selected the ten most abundant transcripts expressed in the leaves (Table 4), and the

**Table 4** Ten most abundant transcripts identified in the *U. humidicola* leaf transcriptome

| Putative gene | E-value | FPKM | UniProtKB |
|---|---|---|---|
| photosystem II reaction center protein M | 7.00E-14 | 12514.83 | sp\|A1E9R2\|PSBM_SORBI |
| uncharacterized ycf68 protein | – | 3330.18 | sp\|P12173 \|YCF68_ORYSJ |
| DNA-binding protein MNB1B | 8.00E-52 | 3004.90 | sp\|P27347\|MNB1B_MAIZE |
| putative uncharacterized protein ART2 | 3.00E-16 | 2583.30 | sp\|Q8TGM7\|ART2_YEAST |
| metallothionein-like protein 1A | 4.00E-19 | 2575.12 | sp\|P0C5B3\|MT1A_ORYSJ |
| thiamine thiazole synthase 2, chloroplastic | 0.0 | 2422.49 | sp\|C5X2M4\|THI42_SORBI |
| protein TIFY 10A | 2.00E-25 | 2167.52 | sp\|Q9LMA8\|TI10A_ARATH |
| cysteine proteinase 1 | 0.0 | 2144.41 | sp\|Q10716\|CYSP1_MAIZE |
| No-hit | – | 2090.78 | – |
| probable polyamine oxidase 2 | 0.0 | 2076.58 | sp\|Q9SKX5\|PAO2_ARATH |

results showed that the common expression pattern was associated with photosynthesis, defense mechanisms and stress responses, which is consistent with results obtained for the *P. maximum* transcriptome [34]. The most represented unigene was Photosystem II reaction center protein M (psbM), which had a FPKM value of 12,514.83. This protein is involved in photosynthesis reactions in photosystem II, which uses light energy to abstract electrons from water to generate a proton gradient that is involved in ATP formation [61]. The second-most abundant unigene was associated with an uncharacterized ycf68 protein. This protein is encoded in the chloroplast and has no known function; however, because this family is exclusively observed in the chloroplasts of phototrophic organisms, this protein might play a role in photosynthesis. In addition to these genes, the DNA-binding protein MNB1B was highly represented. This protein recognizes the AAGG motif at the MNF1 binding site. MNF1 is a nuclear factor that interacts with the promoter region of the phosphoenolpyruvate carboxylase (PEPC) gene, which is involved in the catalysis of the primary fixation of $CO_2$ during C4 photosynthesis [62]. The remaining unigenes were described as participating in the defense mechanisms and stress responses of the plant, including the gene encoding metallothionein-like protein 1A. Metallothioneins are low-molecular-weight cysteine-rich proteins that coordinate metal atoms and are induced under different stress conditions, such as excess heavy metal exposure, heat shock and salt stresses, where these proteins play important roles in maintaining intracellular metal homeostasis, eliminating metal toxicity and protecting against intracellular oxidative damage [63, 64]. The protein thiamine thiazole synthase 2, chloroplastic (THI42) was also abundant, and it has been implicated in the synthesis of the essential cell nutrient thiamine (vitamin B1), or more precisely, the thiazole ring (thiamine precursor) [65]. Moreover, THI42 may play additional roles in adaptation to various stress conditions and DNA damage tolerance. The protein TIFY 10A, which is also known as jasmonate ZIM domain-containing protein 1 (JAZ1), represses the transcription of jasmonate-responsive genes. Thus, this protein is a repressor of jasmonate (JA), an essential phytohormone in plants. The perception of bioactive JAs through the F-box protein coronatine insensitive1 (COI1) leads to the degradation of JAZ1 via the ubiquitin-proteasome pathway, which in turn activates the expression of genes that are involved in plant growth, development, and defense [66]. Cysteine proteases are commonly present in plants and expressed in various organs. These enzymes are involved in digestion, post-translational modification of storage proteins, antibiotic responses and programmed cell death [67]. Specifically, cysteine proteinase 1 is responsible for the degradation of the storage protein zein and may play a role in proteolysis during emergencies. The last unigene among the ten most abundant identified genes was related to the putative polyamine oxidase 2, a flavoenzyme that catalyzes the oxidation of the secondary amino group of polyamines. Polyamine oxidase 2 is located in peroxisomes, which are organelles that are involved in various stress responses [68]. Furthermore, the putative uncharacterized protein ART2 (ribosomal RNA transcript antisense to protein 2) was observed, and among the ten unigenes with higher FPKMs, this transcript did not show similarity to any protein in the Swiss-Prot database. Examination of this unigene for the presence of candidate coding regions revealed a complete ORF. Using this unigene to search against the Phytozome grass data revealed similarity (e-value of 4e-18) with an uncharacterized protein (hypothetical gene #37831442) from *Setaria italica*, *Oryza sativa* and *Brachypodium distachyon*. These unigenes represent genes that have not yet been described, and additional accurate molecular and proteomic analyses are required to validate and determine the functions of these genes.

### Comparison with grass transcriptomes

We searched for orthologs of the *U. humidicola* unigenes by comparing the unigenes of several grass species (species used in the custom databank in the annotation step) to identify unique transcripts for this forage plant. Reciprocal BLAST hits returned 24,133 transcripts with putative orthologous relationships. Compared with each individual grass species in the databank, *U. humidicola* unigenes returned more hits from the *P. maximum* transcriptome (15,427 hits or 43.96 % of the unigenes), whereas *Brachypodium distachyon* returned fewer hits (8,768 hits or 24.98 % of the unigenes) (Additional file 6).

Regarding the unigenes without a relationship to the transcriptomes of other grasses (10,960 unigenes or 31.23 %), 2,231 unigenes were annotated in NCBI nr and 1,010 unigenes were annotated in UniProtKB/Swiss-Prot. We selected these 10,960 unigenes to identify enriched GO terms relative to that of all *U. humidicola* unigenes (Fig. 5). Several classes associated with oxidative stress were identified, including carbon utilization, immune response signaling and catalase activity. Among the 8,729 unigenes that were unique to *U. humidicola* but not annotated, 980 unigenes had ORFs, among which 192 unigene sequences were complete ORFs. This result warrants further investigation of new genes that are potentially exclusive to Koronivia grass.

### Genes involved in flooding stress

Most crops and wild plants cannot tolerate floods and suffer from severe growth reduction or even death under these conditions [69]. However, *U. humidicola* is recognized for its tolerance to poorly drained soils and seasonal flooding as well as infertile acidic soils [8].

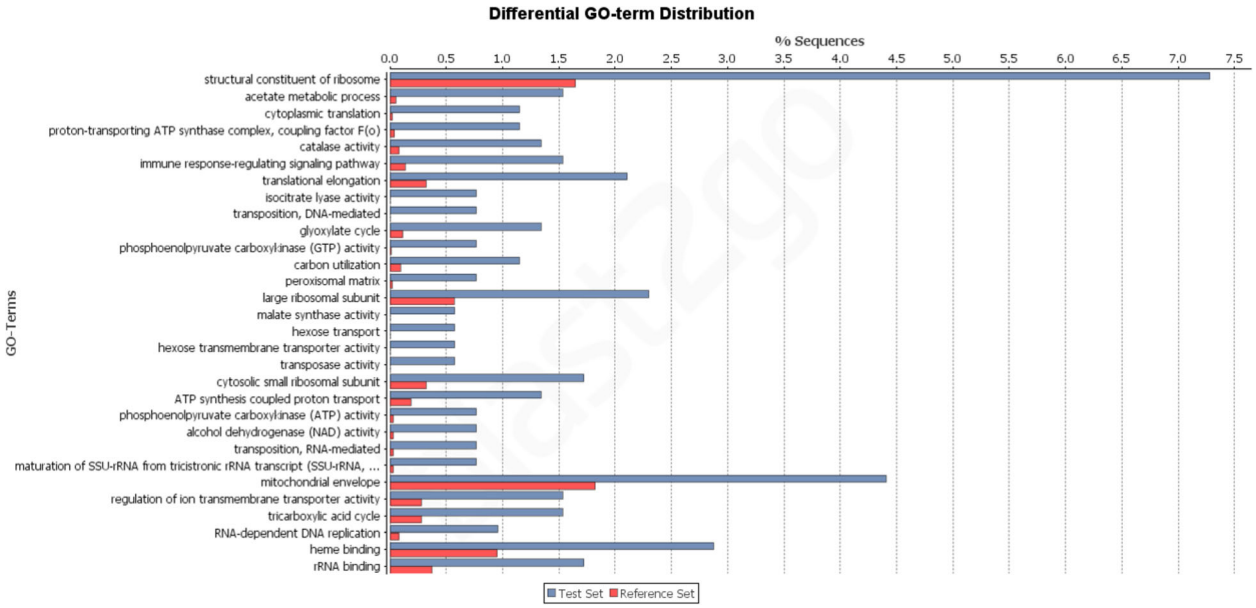


**Fig. 5** Enriched gene ontology (GO) terms for the unique *U. humidicola* unigenes

Although the leaf transcriptome analysis in the present study was not performed under flood stress, we identified genes that are involved in the flooding stress response (Additional file 7), including genes that code for plant hormones, particularly ethylene, ethylene precursors (synthesized through 1-aminocyclopropane-1-carboxylic acid (ACC) synthase [70, 71]), and ethylene response sensors [72]. Here, we identified nine unigene representatives of ACC synthase and four representatives of ethylene response sensors (ERS/ETR1). The ethylene signaling pathway is activated in response to low oxygen stress and initiates and regulates many adaptive molecular, chemical and morphological responses. These responses facilitate the avoidance of anaerobiosis by increasing the availability of oxygen to roots in flooded or waterlogged soil, which may occur through the development of aerenchyma [73, 74]. Eight unigenes were associated with ethylene-insensitive protein 2 (EIN2), a central factor in signaling pathways that is regulated by ethylene and necessary for ethylene-mediated gene regulation. We also identified two ethylene insensitive 3-like 1 proteins that are likely putative transcription factors that function as positive regulators of the ethylene response pathway. In addition, 15 unigenes were related to the ethylene-responsive transcription factor, which is completely dependent on a functional EIN2. We also identified unigenes that were associated with alcohol dehydrogenase (ADH), pyruvate decarboxylase (PDC) [75] and lactate dehydrogenase (LDH) [76] genes, which are expressed in response to hypoxia.

### Genes involved in the C4 photosynthesis pathway

Certain grasses, such as *U. humidicola*, exhibit C4-type photosynthesis, which confers elevated plant productivity, particularly at higher temperatures, in addition to higher water and nitrogen use efficiencies [77]. Analysis of the unigenes revealed 11 different genes that were involved in the C4 pathway (Additional file 8), including phosphoenolpyruvate carboxylase (PEPC) (FPKM value 194.84) and carbonic anhydrase (CA) (FPKM value 398.62), which are enzymes responsible for the initial fixation of $CO_2$ in the C4 mesophyll cell cytosol and subsequent formation of C4 acid oxaloacetate [78]. The C4 acid oxaloacetate is reduced to either malate or aspartate through the enzymes NADP-malate dehydrogenase and aspartate aminotransferase, respectively, which were also identified in the present study. Subsequently, the decarboxylation reaction occurred, which released $CO_2$ in the vicinity of Rubisco, a key step in photosynthetic $CO_2$ assimilation in some C4 plants. The results of the present study demonstrated that the phosphorylation of PEPCK (FPKM 1,116) predominantly occurred in *U. humidicola* in relation to NAD-ME (FPKM 280.01) and NADP-ME (FPKM 7.36). Species such as *Urochloa panicoides* and *Panicum maximum* are considered PEPCK-type C4 plants [34, 77]. PPDK was also representative among the unigenes and had a FPKM value of 1,045.

### Genes involved in cellulose and lignin biosynthesis

Cellulose and lignin are the most abundant polymers in nature. Along with hemicellulose, these polymers are the

major components of the largest source of plant biomass: the lignocellulose wall [79].

Cellulose (b-1,4-glucan) is the predominant polymer of the lignocellulose wall and provides strength and flexibility to plant tissues. This polymer is of great importance to the wood, paper, textile, and chemical industries [80]. Lignin is a complex and heterogeneous mixture of polymers that is primarily derived from three hydroxycinnamyl alcohols (or monolignols): coniferyl alcohol (guaiacyl propanol "G"), sinapyl alcohol (syringyl propanol "S") and p-coumaryl alcohol (p-hydroxyphenyl-propanol "H"). Lignin is the key element that limits cell wall digestibility in ruminants that consume high-forage diets [81]. In general, dicotyledonous angiosperm (hardwood) lignins principally consist of G and S units and traces of H units, whereas lignins from grasses (monocots) incorporate G and S units at comparable levels and present more H units than dicots [82]. The proportion of G:S:H units in the cell wall is an important characteristic for plant breeders and molecular biologists and a determinant of the successful improvement of forages for livestock feeding.

Cellulose is produced from sitosterol-b-glucoside via certain gene families, such as cellulose synthases (CesA) and glycosyl transferases. In the present study, we identified 18 representatives of cellulose synthase A among the *U. humidicola* leaf unigenes (Additional file 9), and our approach also resulted in the identification of other proteins, including sucrose synthase (SuSy) and UTP–glucose-1-phosphate uridylyltransferase, both of which are associated with five unigenes. The most abundant identified transcript was UDP-glycosyltransferase, which had 53 transcript representatives. In addition to other genes in this pathway, seven genes were related to alkaline/neutral invertase. Invertases have been proposed as substitutes for SuSy in nonphotosynthetic cells [83]. Several other proteins have been implemented in cellulose production, such as COBRA-like protein and chitinase-like protein 1, for which we identified eleven and one related unigenes, respectively.

We also mined the current transcriptomic database to obtain unigenes associated with lignin biosynthesis (Additional file 9). We identified enzymes that were putatively involved in phenylpropanoid metabolism and, consequently, monolignol formation. In addition to these enzymes, 12 sequences were associated with chitinases, five sequences were associated with laccases, 37 sequences were associated with peroxidases and nine sequences were associated with dirigent proteins, which are also important for lignification.

### SSR discovery

Next-generation sequencing technology provides access to a wealth of sequence information. Despite the increasing demand for SNP markers for genotyping, SSR markers continue to be important because of they contain large amounts of genetic information that is primarily relevant to parentage composition or forensic studies, where polyallelic variation is useful [84]. SSR markers play an important role in germplasm characterization, linkage and QTL map construction, gene flow and mating system evaluations, and marker-assisted selection [85, 86]. In *U. humidicola*, the currently available SSR markers [11, 21–23] were all developed from enriched genomic libraries.

The unigenes from *U. humidicola* that were obtained in this study represent a valuable resource for SSR mining. Among the 35,093 unigenes, a total of 4,489 putative SSRs were identified in 3,491 unigenes, 566 of which contained more than one SSR and 191 of which were present in compound form (Table 5). The compilation of all SSRs revealed that one SSR on average was identified for every 6.81 kb of the unigenes.

Regarding the unique unigenes containing SSRs of the two sampled genotypes, 516 exclusive unigenes of BRS Tupi included identified SSRs, whereas the BH031 genotype had 190 exclusive unigenes with detected SSRs. The frequency, type and distribution of the 4,489 potential SSRs were also analyzed in the present study. Among the 4,489 SSRs, trinucleotide repeat motifs were the most abundant (3,868, 86.2 %) (Table 5), which is similar to the *P. maximum* transcriptome (86 %) [34] and the *Brachiaria ruziziensis* genome (85.2 %) [87], followed by dinucleotide (363, 8.09 %), tetranucleotide (173, 3.85 %), pentanucleotide (54, 1.20 %) and hexanucleotide (31, 0.69 %) repeat motifs. The total distribution of SSR motifs was similar to that determined for *P. maximum* [34], whereas in *B. ruziziensis* [87], tetranucleotide motifs were more abundant than dinucleotide motifs (Additional file 10). The dinucleotide to hexanucleotide motifs were further analyzed for SSR repeats. Within the SSRs, 231 motif sequence types were identified, of which the di-, tri-, tetra-, penta- and hexa-nucleotide motifs had 12, 60, 94, 44 and 21

**Table 5** Summary of the SSR search results

| Search Item | Numbers |
|---|---|
| Total number of sequences examined | 35,093 |
| Total size of examined sequences (bp) | 30,553,233 |
| Total number of identified SSRs | 4,489 |
| Number of SSR-containing sequences | 3,491 |
| Number of sequences containing more than 1 SSR | 566 |
| Number of SSRs present in compound formation | 191 |
| Dinucleotide | 363 |
| Trinucleotide | 3,868 |
| Tetranucleotide | 173 |
| Pentanucleotide | 54 |
| Hexanucleotide | 31 |

repeats, respectively (Table 6). The most abundant motif detected in the SSRs was the CCG/GGC trinucleotide repeat (1,108, 24.68 %), which is consistent with the results for other grasses, including *P. maximum* [34], *Brachiaria ruziziensis* [87], *Panicum virgatum* [88, 89] and *Hordeum vulgare* [44]. The remaining 258 types of motifs accounted for 5.75 % of the total SSRs (Fig. 6). In summary, we identified a large number of potential functional SSR markers for which primer pairs can be designed and tested to validate the microsatellite loci.

### SNP discovery

A total of 560,298 putative SNP positions were identified in 27,739 different unigenes (79.04 % of the total unigenes), which yielded a density of one SNP position per 119 bp (Table 7). This density was slightly lower than that identified in a previous study for *P. maximum*, which was reported to have a density of 1 SNP per ~90 bp; however, in that study, four genotypes were utilized, thereby increasing the total density of the SNP positions [34]. Among the 27,739 unigenes, 1,427 unigenes were unique to sexual accession BH031 and 4,388 unigenes were unique to cv. BRS Tupi (apomictic cultivar). As expected, transition SNPs (Ts) were more common compared with transversion (Tv) SNPs, resulting in a Ts/Tv ratio of 1.86 (Table 7). Among the transversion variations, C ↔ G was the most highly represented, with 52,161 SNPs, and A ↔ T was the least common, with 45,773 SNPs. The direction and strength of selection maintains transitions over transversions among spontaneous mutations because these parameters generate synonymous mutations in coding sequences [90]. We also identified SNPs in the sequences of unigenes involved in the key pathways described herein. The 59 unigenes related to the flooding stress response contained 1,542 SNPs, whereas the 71 contigs involved in the C4 pathway contained 1,827 SNPs. The 100 unigenes related to cellulose biosynthesis accounted for 2,149 SNPs, whereas the 102 contigs annotated as lignin biosynthesis proteins accounted for 2,067 SNPs. Thus, in the present study, there was a larger number of putative SNP markers in the C4 pathway and lignocellulose biosynthesis pathway compared with the results obtained for *P. maximum* [34], in which 1,159 (C4), 491 (cellulose) and 924 (lignin) different SNP positions were identified. Although all of the predicted molecular markers must be validated to eliminate false positives and sequencing errors, the unigene sequences with SNPs that were identified herein provide an important source of data for the study of *U. humidicola*, for which SNPs have not been previously available.

**Table 6** Length distribution of the SSRs based on the number of repeats

| N. of repeats | Di- | Tri- | Tetra- | Penta- | Hexa- | Total |
|---|---|---|---|---|---|---|
| 4 | - | 2911 | 135 | 44 | 30 | 3120 |
| 5 | - | 718 | 34 | 10 | 0 | 762 |
| 6 | 159 | 168 | 2 | 0 | 0 | 329 |
| 7 | 92 | 62 | 2 | 0 | 0 | 156 |
| 8 | 34 | 8 | 0 | 0 | 0 | 42 |
| 9 | 26 | 1 | 0 | 0 | 0 | 27 |
| 10 | 20 | 0 | 0 | 0 | 1 | 21 |
| 11 | 24 | 0 | 0 | 0 | 0 | 24 |
| ≥12 | 8 | 0 | 0 | 0 | 0 | 8 |

### Differences among genotypes

The samples shared a total of 24,057 unigenes, and the BRS Tupi genotype displayed a larger number of unique transcripts (8,585) compared with that of the BH031 genotype (2,419) (Additional file 11). The difference in the quantity of reads between the two genotypes could have influenced the number of unique transcripts.

At the genomic level [12], BH031 is highly genetically divergent from all other *U. humidicola* germplasm accessions. Moreover, five of the 27 (18.5 %) nuclear microsatellite markers that had previously been developed for cv. BRS Tupi did not amplify a product from BH031 [12], and certain SSR markers that had been developed for BH031 did not amplify a product in BRS Tupi [91] (Additional file 12), indicating a different genomic constitution between the two genotypes. A previous study [11] showed that when these genotypes were crossed, the $F_1$ hybrids presented several meiotic abnormalities and SSR loci that segregated both in disomy and polysomy, indicating an allopolyploid origin of the species, whereas the different genomic constitutions within the species suggested a hybrid origin of the BH031 genotype.

At the transcript level, significant differences were not observed based on Fisher's exact test when searching for enriched GO terms among the analyzed BRS Tupi and BH031 plants. We identified 3,202 ORFs that were unique to BRS Tupi, 859 that were unique to BH031 and 18,520 that were shared between the two genotypes. To validate the expression results, three different plants from each genotype were analyzed via qPCR.

### Validation of the differential expression of transcripts among the sexual and apomictic genotypes

Seven reference genes were selected, and primer pairs were designed based on the *U. humidicola* transcriptome sequences (Additional file 13). At the amplification stage, the Uhum_Ea1 and Uhum_18S genes did not amplify or provide different DNA band intensities between the two evaluated genotypes and were discarded. Five primers pairs were then evaluated for qPCR: Uhum_Actin, Uhum_Ubiquitin, Uhum_GAPDH1, Uhum_GAPDH3 and Uhum_Ubiquitin-40S. In the amplification efficiency test, Uhum_GAPDH3 did not provide an E value from 90-

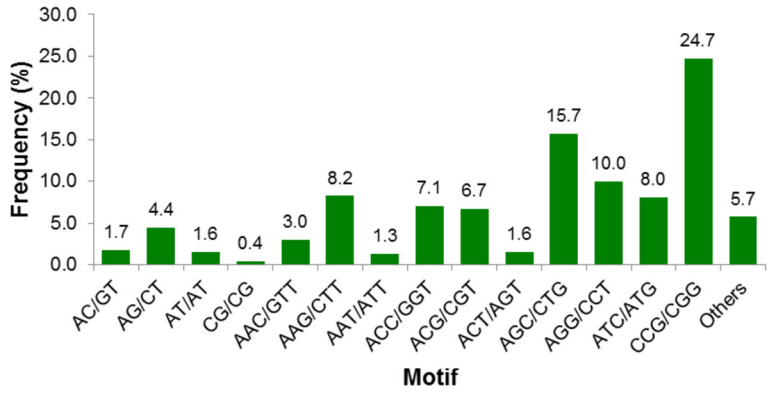


**Fig. 6** Frequency distribution of SSRs based on the motif sequence types

110 %. Among the other four primer pairs, the M values were all < 0.5 and the CV values were all < 0.25 (Additional file 14). The Uhum_Ubiquitin and Uhum_Ubiquitin-40s genes presented the lowest M and CV values and were used as reference genes in the present study. Ubiquitin-related genes are usually reported as good reference genes in grasses [92, 93], and the genes selected herein are the first reference genes to be validated for gene expression studies of *U. humidicola* and can be used in other *Urochloa* species (upon validation) as previously described [47] for *U. brizantha*.

Primer pairs were designed for ten differentially expressed transcripts among the two genotypes and ten and eleven BH031 and BRS Tupi unique transcripts, respectively (Additional file 15). All of the primer pairs were evaluated via PCR using genomic DNA as a template. From the differentially expressed category, nine of the ten pairs successfully provided amplicons. Among the ten primer pairs of the BH031 unique transcripts, three primer pairs amplified a product in both genotypes, and among the BRS Tupi unique transcripts, only one primer pair amplified a product in the BH031 genotype (Additional files 16 and 17). Genomic microsatellite primer pairs that were derived from one genotype and analyzed in the other genotype did not amplify a product, indicating that these genotypes were genetically divergent (Additional file 12) as previously observed [12, 91].

After the amplification efficiency analysis of the cDNA, seven primer pairs were selected for the expression analysis: four primer pairs from differentially expressed genes and three primer pairs from unique BH031 transcripts. The evaluated transcripts from the differentially expressed sequences among the analyzed genotypes displayed similarities to the following proteins: triose phosphate/phosphate translocator (TPT; carbon and phosphoenol-pyruvate transport [94, 95]), cysteine proteinase 2 (CYSP2; degradation and mobilization of storage proteins in the endosperm [96, 97]), sodium/metabolite cotransporter BASS2 (sodium-dependent pyruvate uptake activity in the plastids [98]) and thiamine thiazole synthase 2 (THI42 [99]). Among the BH031 unique transcripts, two sequences exhibited similarities to premnaspirodiene oxygenase (C7D55; involved in the biosynthesis of solavetivone, a potent antifungal phytoalexin [100]) and calmodulin-binding transcription factor (CMTA1; regulates transcriptional activity in response to calcium signals [101]) proteins, whereas the third sequence showed no similarity to any protein present in the NCBI nr database (no-hit).

Among the differentially expressed sequences, the TPT transcript displayed the higher difference in expression among the analyzed genotypes; however, the sequence showed greater expression in the BRS-Tupi genotypes (Fig. 7a). Although the CYSP2, BASS2 and THI42 transcripts were more highly expressed in the BH031 genotype in the *in silico* analysis, significant differences were not observed in the expression level of these sequences between BH031 and BRS Tupi plants in the qPCR assays (Figs. 7b, c and d). Among the BH031 unique transcripts, all three transcripts were highly expressed in the three plants (Fig. 7e, f and g). The

**Table 7** Summary of the putative SNPs identified using the CLC Genomics Workbench

| | |
|---|---|
| Number of unigenes | 35,093 |
| Total bases | 66,854,141 |
| Number of SNPs | 560,298 |
| SNP frequency | 1 per 119 bp |
| *Transitions* | 358,916 |
| A ↔ G | 179,262 |
| C ↔ T | 179,654 |
| *Transversions* | 192,409 |
| A ↔ C | 46,978 |
| A ↔ T | 45,773 |
| C ↔ G | 52,161 |
| G ↔ T | 47,497 |

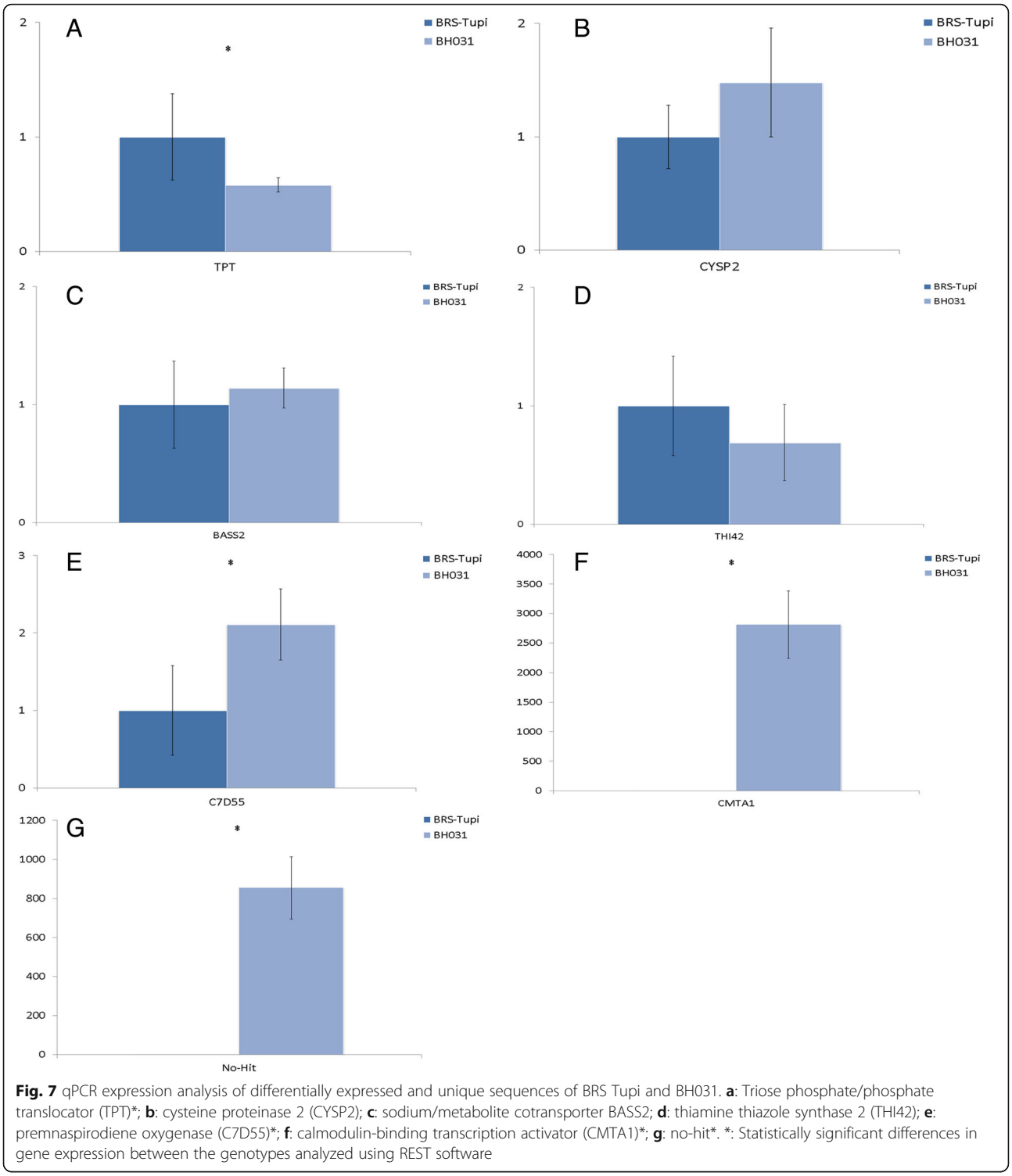


**Fig. 7** qPCR expression analysis of differentially expressed and unique sequences of BRS Tupi and BH031. **a**: Triose phosphate/phosphate translocator (TPT)*; **b**: cysteine proteinase 2 (CYSP2); **c**: sodium/metabolite cotransporter BASS2; **d**: thiamine thiazole synthase 2 (THI42); **e**: premnaspirodiene oxygenase (C7D55)*; **f**: calmodulin-binding transcription activator (CMTA1)*; **g**: no-hit*. *: Statistically significant differences in gene expression between the genotypes analyzed using REST software

sequence that was similar to the C7D55 protein presented a 2-fold higher expression level in BH031 plants than in BRS-Tupi plants. The CMTA1 transcript was the most differentially expressed among the genotypes and exhibited nearly 3,000-fold higher expression level in BH031 plants. The no-hit sequence was also highly expressed in the sexual genotypes and showed a 900-fold higher expression level than that in the apomictic genotypes. In general, the results indicated that *in silico* analyses are useful for identifying differentially expressed and unique genes. Certain genes showed divergence among the RT-PCR and quantitative RNA-seq analyses,

which could be explained by a lack of sequencing depth and suggests that a new analysis using biological replicates and shorter reads *per* library could provide better results for the *in silico* quantification of mRNA expression.

A transcript that presented similarity to a CMTA protein was apparently silenced in the apomictic genotypes because a DNA coding region was observed but genic expression was not observed; therefore, we decided to verify the presence of other sequences in the transcriptome that showed similarity to CMTA proteins. *CMTA* genes respond differentially and rapidly to hormonal stimuli and environmental stresses [102, 103]. In the family Panicoideae, the number of *CMTA* genes ranges from six (*S. viridis*) to 14 (*P. virgatum*) (Phytozome database, as of June 2016). Another eight transcripts were annotated as CMTA proteins: one for CMTA1, four for CMTA2, two for CMTA4 and one for CMTA6 (Additional file 18). Read mapping statistics showed that two sequences similar to CMTA2 proteins and one sequence similar to CMTA6 proteins had comparable numbers of reads for both genotypes. In contrast, the BRS-Tupi plant presented a larger number of reads for one transcript that presented similarity to a CMTA1 protein and two transcripts that presented similarity to the CMTA4 protein. The sexual genotype BH031 presented a larger number of reads for the other transcripts similar to the CMTA2 protein, and one of these transcripts only presented reads in this genotype.

Our results and the SSR amplification profiles suggest that one of these genotypes likely presents one or more genomic regions that are absent in the other genotype. In addition, certain sequences are silenced in the BRS-Tupi genotypes but transcribed in the BH031 plants. Because the BH031 accession is the only known sexual genotype, we hypothesize that BH031 originated from a cross of *U. humidicola* and an unknown parent. Interspecific hybridization causes considerable changes in gene expression [104], including a loss of paternal imprinting in the hybrids [105]. The BH031 accession would represent a hybrid that has a genome derived from a combination of the genomes of both parents. This scenario would explain the amplification failure of genomic SSRs and transcribed sequences in both genotypes as well as the transcriptional activation of sequences in the BH031 genotype that are silenced in the BRS-Tupi genotype. Further cytogenetic studies are needed to confirm this hypothesis.

### RNA-seq for *U. humidicola* breeding

The domestication and breeding of tropical forage grasses were initiated some decades ago. Similar to other perennial species [55], Koronivia grass remains in the early stages of domestication, which increases the difficulty of identifying concentrated gene groups that play specific roles in important agronomic characteristics. The Koronivia grass breeding program has applied selection methods for many decades, and hybrid progenies were obtained in the early 2000s. Koronivia grass is a perennial crop and necessitate animal trials to release breeding materials, leading to a new cultivar every 10–15 years [1]. In contrast, many agricultural annual crops produce 2–3 generations each year and release cultivars annually.

Next-generation sequencing (NGS)-derived methodologies, such as RNA-seq, have provided rapid advances in genomic data availability. The present study represents an advancement in the description of the *U. humidicola* transcriptome and provides abundant available data via the 2,237 ESTs that were deposited in the NCBI database in March 2015 and more than 35,095 unigenes that were identified via leaf RNA-seq, which included 5,324 unigenes assigned to 327 known metabolic pathways. Moreover, this methodology has facilitated the identification of novel transcripts and new functional markers and improved the SSR database available for this species [11, 22–24].

The high genetic variability available for this species has been demonstrated based on SSR polymorphisms [12]; however, SNP markers are the most abundant type of DNA polymorphism in genomic sequences, and major phenotypic variations have been assigned to this class of markers [106]. Thus, the combination of the RNA-seq approach and SNP identification is ideal for the development of new markers in candidate genes for genetic breeding, and along with conventional breeding, this technique can enhance *U. humidicola* domestication.

## Conclusions

The RNA-seq approach has improved our current knowledge of the transcriptional patterns in the leaves of *U. humidicola* and facilitated the identification of potentially novel genes. The comparison between unigenes that are exclusive to the sexual and apomictic genotypes analyzed herein will allow for the identification and characterization of genes that might be related to aposporic apomixis in this species; however, further evaluation of the reproductive tissues is required. In this study, 4,489 new EST-SSRs and 560,298 new SNP markers that may be associated with important genes for the studied species were identified. Specific SNPs in the sequences of the unigenes involved in the flooding stress response, the C4 pathway and the biosynthesis of cellulose and lignin were identified. The latter could be used to improve forage quality, which is one of the main issues associated with *U. humidicola*. These results provide valuable information related to the genomic resources of *U. humidicola* and other brachiaria grasses for breeding programs. Moreover, important differences between the transcriptomes of the genotypes corroborate genomic observations, which will

facilitate investigations of the potential hybrid origin of the sexual BH031 accession. Finally, the *in silico* analysis method was useful for identifying differentially expressed and unique genes in both genotypes.

## Additional files

**Additional file 1:** Comparison of the *U. humidicola* assembly metrics for the Phytozome transcriptomes. (XLSX 9 kb)

**Additional file 2:** Species top hits in the NCBI nr database. The five species with greater numbers of top hits in the NCBI nr database compared with the *U. humidicola* unigenes via BLASTX. (PNG 150 kb)

**Additional file 3:** KEGG glycolysis/gluconeogenesis pathway. The genes that were present in the *U. humidicola* transcriptome are indicated in blue. (PNG 702 kb)

**Additional file 4:** KEGG plant hormone signal transduction pathway. The genes that were present in the *U. humidicola* transcriptome are indicated in blue. (PNG 771 kb)

**Additional file 5:** Open reading frames (ORFs) in the *U. humidicola* transcriptome. Description of the types of ORFs identified in the transcript and unigene datasets of the *U. humidicola* transcriptome. (XLSX 9 kb)

**Additional file 6:** Reciprocal BLAST hit of the assembled unigenes from the *U. humidicola* transcriptome against other grass transcriptomes. Homology search by BLASTn with a cutoff value of 1e-10. (PNG 271 kb)

**Additional file 7:** List of flooding stress-related genes identified among *U. humidicola* unigenes. (XLSX 9 kb)

**Additional file 8:** List of genes composing the C4 photosynthetic pathway among the *U. humidicola* unigenes. (XLSX 8 kb)

**Additional file 9:** List of genes composing the cellulose and lignin pathways among the *U. humidicola* unigenes. (XLSX 10 kb)

**Additional file 10:** Comparison of the SSRs from *U. humidicola*, *Panicum maximum* and *Brachiaria ruziziensis*. (XLSX 9 kb)

**Additional file 11:** Venn diagram representing the shared and unique unigenes in the *U. humidicola* transcriptome. (PNG 549 kb)

**Additional file 12:** Microsatellite amplification profile for the BH031 and cv. BRS Tupi genotypes. Microsatellite amplification profile showing the lack of amplification in the BH031 and cv. BRS Tupi genotypes, which is dependent on the genotype from which the SSR was developed. Loci BhUNICAMP084 (A) and BhUNICAMP065 (B) developed from BH031-enriched libraries [106] and showing the amplification of the SSR markers in BH031 (P1) but not in cv. BRS Tupi (P2); and loci BhUNICAMP082 (C) and BhUNICAMP100 (D) developed from cv. BRS Tupi-enriched libraries [62] and showing the amplification of these SSR markers in cv. BRS Tupi (P2) but not in BH031 (P1). The remaining individuals shown in the polyacrylamide gels are $F_1$ hybrids from the cross BH031 x cv. BRS Tupi. (TIF 11136 kb)

**Additional file 13:** Primer sequences and amplicons of the seven candidate reference genes evaluated in this study. (XLSX 10 kb)

**Additional file 14:** Gene expression stability values (M) and coefficient of variance (CV) of the candidate reference genes. (XLSX 9 kb)

**Additional file 15:** Quantitative RT-PCR primer sequences. (XLSX 13 kb)

**Additional file 16:** Amplification of primer pairs using genomic DNA. BH031 (first sample) and BRS Tupi (second sample). A: Sequences containing reads from both genotypes, B: sequences containing reads from BRS Tupi only, C: sequences containing reads from BH031 only. All of the primer pairs are described in Additional file 15. (PNG 1555 kb)

**Additional file 17:** Amplification success of the 31 primer pairs for the qPCR assays for both DNA and RNA of genotypes BRS-Tupi and BH031. (XLSX 10 kb)

**Additional file 18:** Number of reads from the BRS-Tupi and BH031 accessions that presented similarity to CMTA proteins. (XLSX 9 kb)

### Abbreviations

4CL: 4-Hydroxycinnamoyl CoA ligase; ACC synthase: 1-Aminocyclopropane-1-carboxylic acid synthase; ADH: Alcohol dehydrogenase; ART2: Ribosomal RNA transcript antisense to protein 2; BASS2: Sodium/metabolite cotransporter; C3H: P-coumarate 3-hydroxylase; C7D55: Premnaspirodiene; CA: Carbonic anhydrase; CesA: Cellulose synthases; CMTA1: Calmodulin-binding transcription factor; CNPq: Brazilian national council for scientific and technological development; COB: Cobra; COI1: Coronatine Insensitive1; CTL1/POM1: Chitinase-like protein 1; CYP450: Cytochrome P450; CYSP2: Cysteine proteinase 2; EIN2: Ethylene-insensitive protein 2; Embrapa: Brazilian agricultural research Corporation; ERS/ETR1: Ethylene-response sensors; ESTs: Expressed sequence tags; F5H: Ferulate 5-hydroxylase; FAPESP: State of São Paulo Research Foundation; FPKM: Fragments per kilobase of transcript per million mapped reads; G: Coniferyl alcohol (guaiacyl propanol); GO: Gene ontology; H: p-Coumaryl alcohol (p-hydroxyphenylpropanol); JA: Jasmonate; JAZ1: Jasmonate ZIM domain-containing protein 1; KAAS: KEGG automatic annotation server; KEGG: Kyoto encyclopedia of genes and genomes database; KO: KEGG orthology; KOB1: Kobito1; KOR: Korrigan; LDH: Lactate dehydrogenase; MISA: MIcroSAtellite script; NAD-ME: NAD-malic enzyme; NADP-ME: NADP-malic enzyme; NCBI: National center of biotechnology information; NGS: Next-generation sequencing; Nr: National center for biotechnology information non-redundant protein database; OLB: Off-Line basecaller software; ORFs: Open reading frames; PCR: Polymerase chain reaction; PDC: Pyruvate decarboxylase; PE: Paired end; PEP: Phosphoenolpyruvate; PEPC: Phosphoenolpyruvate carboxylase; PEPCK: Phosphoenolpyruvate carboxykinase; PPDK: Pyruvate phosphate dikinase; PPR: Pentatricopeptide repeat; psbM: Photosystem II reaction center protein M; qPCR: Quantitative RT-PCR; RBH: Reciprocal BLAST hit; RSEM: RNA-seq by Expectation Maximization; RT-PCR: Reverse transcription polymerase chain reaction; S: Sinapyl alcohol (syringyl propanol); SAD: Sinapyl alcohol dehydrogenase; SNPs: Single nucleotide polymorphisms; SRA: Short read archive; SSR: Single sequence repeats; SuSy: Sucrose synthase; THI42: Thiamine thiazole synthase 2, chloroplastic; TPT: Triose phosphate/phosphate translocator; Ts: Transition SNPs; Tv: Transversion SNPs; UNICAMP: University of Campinas

### Funding

This work was financially supported by grants from the Brazilian National Council for Scientific and Technological Development (CNPq, grant numbers 478262/2004-3, 502336/2005-6 and 482458/2007-0), Computational Biology Program from Coordination of Superior Level Staff Improvement (CAPES 15/2013) and the State of São Paulo Research Foundation (FAPESP, 2008/52197-4). The authors would like to thank FAPESP for providing a graduate fellowship (2013/14903-2) to FAO and a post-graduate fellowship (2013/20447-0) to GTS. The authors would also like to thank the CNPq for providing a post-graduate fellowship (150719/2015-9) to CCS and a research fellowship (307430/2007-3) to APS.

### Availability of data and materials

The reads datasets supporting the results of this article are deposited in the NCBI Short Read Archive (SRA) under accession number SRP065020, at the following link: https://www.ncbi.nlm.nih.gov/sra/?term=SRP065020.

### Authors' contributions

Conceived and designed the experiments: BBZV GTS CBV APS. RNA extraction, cDNA library generation and sequencing: GTS BBZV. Performed the experiments: FAO BBZV GTS CCS. Analyzed the data: GTS FAO BBZV CCS. Contributed material/reagents/analysis tools: BBZV FAO GTS CCS CBV APS. Drafted the manuscript: BBZV FAO GTS CCS CBV APS.

### Competing interests

This work is a collaborative research project that was developed by researchers from UNICAMP (Brazil) and EMBRAPA (Brazil), and the authors declare that they have no competing interests.

### Consent for publication

Not applicable.

### Ethics approval and consent to participate

All of the necessary permits for the field studies were obtained within Embrapa. The field studies did not involve native, endangered or protected species.

**Author details**
[1]Embrapa Pecuária Sudeste, São Carlos, SP, Brazil. [2]Center for Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Campinas, SP, Brazil. [3]Embrapa Gado de Corte, Campo Grande, MS, Brazil. [4]Department of Plant Biology, Biology Institute, UNICAMP, Campinas, SP, Brazil. [5]Present Address: Department of Biochemistry, Center of Biological Sciences, Federal University of Santa Catarina, Florianopolis, SC, Brazil.

Received: 9 April 2016 Accepted: 5 November 2016
Published online: 11 November 2016

**References**

1. do Valle CB, Simioni C, Resende RMS, Jank L, Chiari L. Melhoramento genético de *Brachiaria*. In: Resende RMS, do Valle CB, Jank L, editors. Melhoramento de forrageiras tropicais. 1st ed. Campo Grande: Embrapa; 2008. p. 13–53.
2. Ministério da Agricultura; 2014. http://www.agricultura.gov.br/animal/especies/bovinos-e-bubalinos. Accessed 13 November 2014.
3. Anualpec. Anuario da pecuária brasileira. São Paulo: Informa Economics FNP; 2008.
4. Valle CB, Jank L, Resende RMS. O melhoramento de forrageiras tropicais no Brasil. Revista Ceres. 2009;56:460–72.
5. Boldrini KR, Pagliarini MS, Valle CB. Cell fusion and cytomixis during microsporogenesis in *Brachiaria humidicola* (Poaceae). South Afr J Bot. 2006; 72:478–81.
6. Watson L, Dallwitz MJ; 1992. Onwards. The grass genera of the world: descriptions, illustrations, identification, and information retrieval; including synonyms, morphology, anatomy, physiology, phytochemistry, cytology, classification, pathogens, world and local distribution, and references. http://delta-intkey.com. Accessed 25 September 2015.
7. Morrone O, Zuloaga FO. Revisión de las especies sudamericanas nativas y introducidas de los géneros *Brachiaria* y *Urochloa* (Poaceae: Panicoideae: Paniceae). Darwiniana. 1992;31:43–109.
8. Keller-Grein G, Maass BL, Hanson J. Natural variation in *Brachiaria* and existing germplasm collections. In: Miles JW, Maass BL, Valle CB, editors. *Brachiaria*: biology, agronomy and improvement. Cali: Embrapa/CIAT; 1996. p. 16–42.
9. Assis GML, dos Santos CF, Flores PS, do Valle CB. Genetic divergence among Brachiara humidicola (Rendle) Schweick hybrids evaluated in the Western Brazilian Amazon. Crop Breed Appl Biotechnol. 2014;14:224–31.
10. Figueiredo UJ, Nunes JAR, Valle CB. Estimation of genetic parameters and selection of *Brachiaria humidicola* progenies using a selection index. Crop Breed Appl Biotechnol. 2012;12:237–44.
11. Vigna BB, Santos JC, Jungmann L, do Valle CB, Mollinari M, Pastina MM, et al. Evidence of allopolyploidy in *Urochloa humidicola* based on cytological analysis and genetic linkage mapping. PLoS One. 2016;11:e0153764.
12. Jungmann L, Vigna BB, Boldrini KR, Sousa AC, do Valle CB, Resende RM, et al. Genetic diversity and population structure analysis of the tropical pasture grass *Brachiaria humidicola* based on microsatellites, cytogenetics, morphological traits, and geographical origin. Genome. 2010;53:698–709.
13. Boldrini KR, de Victor Adamowski E, Message H, Calisto V, Pagliarini MS, do Valle CB. Meiotic behavior as a selection tool in the breeding of *Brachiaria humidicola* (Poaceae). Euphytica. 2011;182:317–24.
14. Adamowski EV, Boldrini KR, Pagliarini MS, do Valle CB. Abnormal cytokinesis in microsporogenesis of *Brachiaria humidicola* (Poaceae: Paniceae). Genet Mol Res. 2007;6:616–21.
15. Boldrini KR, Pagliarini MS, Valle CB. Meiotic behavior of a nonaploid accession endorses x = 6 for Brachiaria humidicola (Poaceae). Genet Mol Res. 2009;8:1444–50.
16. Ishigaki G, Gondo T, Ebina M, Suenaga K, Akashi R. Estimation of genome size in *Brachiaria* species. Grassl Sci. 2010;56:240–2.
17. Jank L, Valle C, Resende R. Breeding tropical forages. Crop Breed Appl Biotechnol. 2011;11:27–34.
18. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods. 2008;5:16–8.
19. Ansorge WJ. Next-generation DNA, sequencing techniques. New Biotechnol. 2009;25:195–203.
20. Ha K, Lim CJ, Kim S, Choe JK, Jo S-H, Baek N, et al. High-throughput sequencing and de novo assembly of *Brassica oleracea* var. *Capitata* L. for transcriptome analysis. PLoS One. 2014;9:e92087.
21. Jungmann L, Vigna BBZ, Paiva J, Sousa ACB, do Valle CB, Laborda PR. Development of microsatellite markers for *Brachiaria humidicola* (Rendle) Schweick. Conserv Genet Resour. 2009;1:475–9.
22. Vigna BB, Alleoni GC, Jungmann L, do Valle CB, de Souza AP. New microsatellite markers developed from *Urochloa humidicola* (Poaceae) and cross amplification in different *Urochloa* species. BMC Res Notes. 2011;4:523.
23. Santos JC, Barreto MA, Oliveira FA, Vigna BB, Souza AP. Microsatellite markers for *Urochloa humidicola* (Poaceae) and their transferability to other *Urochloa* species. BMC Res Notes. 2015;8:83.
24. EMBRAPA. Embrapa produtos e mercado; 2012. http://ainfo.cnptia.embrapa.br/digital/bitstream/item/77436/1/Folder-Tupi-Junho2012-CV.pdf. Accessed 1 Apr 2014.
25 Oñate-Sanchez L, Vicente-Carbajosa J. DNA-free RNA isolation protocols for *Arabidopsis thaliana*, including seeds and siliques. BMC Res Notes. 2008;1:93.
26 Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 1987;19:11–5.
27 Patel RK, Jain M. NGS QC toolkit: a toolkit for quality control of next generation sequencing data. Plos One. 2012;7:e30619.
28 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
29 Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010;95:315–27.
30 Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
31 Magrane M, UniProt consortium. UniProt Knowledgebase: a hub of integrated protein data. Database. 2011;2011:bar009. doi:10.1093/database/bar009.
32 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
33 Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40:D1178–86.
34 Toledo-Silva G, Cardoso-Silva CB, Jank L, Souza AP. *De novo* transcriptome assembly for the tropical grass *panicum maximum* Jacq. PLoS One. 2013;8:e70781.
35 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
36 Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–6.
37 Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, et al. WEGO: a web tool for plotting GO annotations. Nucleic Acids Res. 2006;34:W293–7.
38 Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
39 Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35:W182–5.
40 Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, et al. The Pfam protein families database. Nucleic Acids Res. 2004;32:D138–41.
41 Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. BioInformatics. 2001;17:847–8.
42 Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
43 Oliveros JC. Venny v.2.0.2 computational genomics; 2015. http://bioinfogp.cnb.csic.es/tools/venny/index.html. Accessed 21 October 2015.
44 Thiel T, Michalek W, Varshney RK, Graner A. Exploiting est databases for the development and characterization of gene-derived ssr-markers in barley (Hordeum vulgare l.). Theor Appl Genet. 2003;106:411–22.
45 Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. 2007;35:W71–4.
46 Hong SY, Seo PJ, Yang MS, Xiang F, Park CM. Exploring valid reference genes for gene expression studies in *Brachypodium distachyon* by real-time PCR. BMC Plant Biol. 2008;8:112.
47 Silveira ED, Alves-Ferreira M, Guimarães LA, da Silva FR, Carneiro VT. Selection of reference genes for quantitative real-time PCR expression studies in the apomictic and sexual grass *Brachiaria brizantha*. BMC Plant Biol. 2009;9:84.
48 Simon B, Conner JA, Ozias-Akins P. Selection and validation of reference genes for gene expression analysis in apomictic and sexual *Cenchrus ciliaris*. BMC Res Notes. 2013;6:397.

49 Gimeno J, Eattock N, Van Deynze A, Blumwald E. Selection and validation of reference genes for Gene expression analysis in switchgrass (*panicum virgatum*) using quantitative real-Time RT-PCR. Plos One. 2014;9:e91474.
50 Pfaffl MW, Horgan GW, Dempfle L. Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Res. 2002;30:e36.
51 Liu M, Qiao G, Jiang J, Yang H, Xie L, Xie J, et al. Transcriptome sequencing and *de novo* analysis for ma bamboo (*Dendrocalamus latiflorus Munro*) using the Illumina platform. Plos One. 2012;7:e46766.
52 Ge X, Chen H, Wang H, Shi A, Liu K. *De novo* assembly and annotation of *Salvia splendens* Transcriptome using the Illumina platform. Plos One. 2014;9:e87693.
53 Peng Y, Gao X, Li R, Cao G. Transcriptome sequencing and *de novo* analysis of *Youngia japonica* using the Illumina platform. Plos One. 2014;9:e90636.
54 Wei L, Li S, Liu S, He A, Wang D, Wang J, et al. Transcriptome analysis of *Houttuynia cordata* Thunb. By Illumina paired-end RNA sequencing and SSR marker discovery. Plos One. 2014;9:e84105.
55 Mantello CC, Cardoso-Silva CB, da Silva CC, de Souza LM, Scaloppi Junior EJ, de Souza GP, et al. De novo assembly and Transcriptome analysis of the rubber tree (*hevea brasiliensis*) and SNP Markers development for rubber biosynthesis pathways. Plos One. 2014;9:e102665.
56 Bennetzen JL, Freeling M. The unified grass genome: synergy in Synteny. Genome Res. 1997;7:301–6.
57 Small ID, Peeters N. The PPR motif - a TPR-related motif prevalent in plant organellar proteins. J Trends Biochem Sci. 2000;25:46–7.
58 Lurin C, Andrés C, Aubourg S, Bellaoui M, Bitton F, Bruyère C, et al. Genome-wide analysis of *Arabidopsis* Pentatricopeptide repeat proteins reveals their essential role in organelle Biogenesis. Plant Cell. 2004;16:2089–103.
59 van der Biezen EA, Jones JDG. The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. Curr Biol. 1998;8:R226–7.
60 Nelson D, Werck-Reichhart D. A P450-centric view of plant evolution. Plant J. 2011;66:194–211.
61 Kawakami K, Umena Y, Iwai M, Kawabata Y, Ikeuchi M, Kamiya N, et al. Roles of PsbI and PsbM in photosystem II dimer formation and stability studied by deletion mutagenesis and X-ray crystallography. Biochim Biophys Acta. 2011;1807(3):319–25.
62 Yanagisawa S, Izui K. Molecular cloning of two DNA-binding proteins of maize that are structurally different but interact with the same sequence motif. J Biol Chem. 1993;268:16028–36.
63 Cobbett C, Goldsbrough P. Phytochelatins and metallothioneins: roles in heavy metal detoxification and homeostasis. Annu Rev Plant Biol. 2002; 53:159–82.
64 Zhou G, Xu Y, Li J, Yang L, Liu JY. Molecular analyses of the metallothionein gene family in rice (*Oryza sativa* L.). J Biochem Mol Biol. 2006;39:595–606.
65 Dorrestein PC, Zhai H, McLafferty FW, Begley TP. The biosynthesis of the thiazole phosphate moiety of thiamin: the sulfur transfer mediated by the sulfur carrier protein ThiS. Chem Biol. 2004;11:1373–81.
66 Chung HS, Howe GA. A critical role for the TIFY motif in repression of jasmonate signaling by a stabilized splice variant of the JASMONATE ZIM-domain protein JAZ10 in *Arabidopsis*. Plant Cell. 2009;21:131–45.
67 Kiyosaki T, Asakura T, Matsumoto I, Tamura T, Terauchi K, Funaki J. Wheat cysteine proteases triticain α, β and γ exhibit mutually distinct responses to gibberellin in germinating seeds. J Plant Physiol. 2009;166:101–6.
68 Moschou PN, Sanmartin M, Andriopoulou AH, Rojo E, Sanchez-Serrano JJ, Roubelakis-Angelakis KA. Bridging the gap between plant and mammalian polyamine catabolism: a novel peroxisomal polyamine oxidase responsible for a full back-conversion pathway in *Arabidopsis*. Plant Physiol. 2008;147:1845–57.
69 Colmer TD, Voesenek LACJ. Flooding tolerance : suites of plant traits in variable environments. Funct Plant Biol. 2009;36:665–81.
70 Grichko VP, Glick BR. Ethylene and flooding stress in plants. Plant Physiol Biochem. 2001;39:1–9.
71 Shiu OY, Oetiker JH, Yip WK, Yang SF. The promoter of LE-ACS7, an early flooding-induced 1-aminocyclopropane-1-carboxylate synthase gene of the tomato, is tagged by a Sol3 transposon. Proc Natl Acad Sci U S A. 1998;95:10334–9.
72 Voesenek LACJ, Rijnders JHGM, Peeters AJM, Van de Steeg HM, de Kroon H. Plant hormones regulate fast shoot elongation under water: from genes to communities. Ecol. 2004;85:16–27.
73 Irfan M, Hayat S, Hayat Q, Afroz S, Ahmad A. Physiological and biochemical changes in plants under waterlogging. Protoplasma. 2010;241:3–17.
74 Sairam RK, Kumutha D, Ezhilmathi K, Deshmukh PS, Srivastava GC. Physiology and biochemistry of waterlogging tolerance in plants. Biol Plant. 2008;52:401–12.
75 Christianson JA, Llewellyn DJ, Dennis ES, Wilson IW. Comparisons of early transcriptome responses to low-oxygen environments in three dicotyledonous plant species. Plant Signal Behav. 2010;5:1006–9.
76 Drew MC. Oxygen deficiency and root metabolism: injury and acclimation under hypoxia and anoxia. Annu Rev Plant Physiol Plant Mol Biol. 1997;48:223–50.
77 Leegood RC. Strategies for engineering C(4) photosynthesis. J Plant Physiol. 2013;170:378–88.
78 Matsuoka M, Furbank RT, Fukayama H, Miyao M. Molecular engineering of C4 photosynthesis. Annu Rev Plant Physiol Plant Mol Biol. 2001;52:297–314.
79 Pérez J, Muñoz-Dorado J, de la Rubia T, Martínez J. Biodegradation and biological treatments of cellulose, hemicellulose and lignin: an overview. Int Microbiol. 2002;5:53–63.
80 Peng L, Kawagoe Y, Hogan P, Delmer D. Sitosterol-beta-glucoside as primer for cellulose synthesis in plants. Science. 2002;295:147–50.
81 Jung HG, Allen MS. Characteristics of plant cell walls affecting intake and digestibility of forages by ruminants. J Anim Sci. 1995;73:2774–90.
82 Boerjan W, Ralph J, Baucher M. Lignin biosynthesis. Annu Rev Plant Biol. 2003;54:519–46.
83 Barratt DH, Derbyshire P, Findlay K, Pike M, Wellner N, Lunn J. Normal growth of *Arabidopsis* requires cytosolic invertase but not sucrose synthase. Proc Natl Acad Sci U S A. 2009;106:13124–9.
84 Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, Seeb LW. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Mol Ecol Res. 2011;11:1–8.
85 Kumar P, Gupta VK, Misra AK, Modi DR, Pandey BK. Potential of Molecular Markers in Plant Biotechnology. Plant Omics Journal. 2009;2:141–62.
86 Powell W, Machray G, Provan J. Polymorphism revealed by simple sequence repeats. Trends Plant Sci. 1996;1:215–22.
87 Silva PI, Martins AM, Gouvea EG, Pessoa-Filho M, Ferreira ME. Development and validation of microsatellite markers for *Brachiaria ruziziensis* obtained by partial genome assembly of Illumina single-end reads. BMC Genomics. 2013;14:17.
88 Wang Y, Zeng X, Iyer NJ, Bryant DW, Mockler TC, Mahalingam R. Exploring the switchgrass transcriptome using second-generation sequencing technology. Plos One. 2012;7:e34225.
89 Sharma MK, Sharma R, Cao P, Jenkins J, Bartley LE, Qualls M, et al. A genome-wide survey of switchgrass genome structure and organization. PLoS One. 2012;7:e33892.
90 Wakeley J. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Trends Ecol Evol. 1996;11:158–62.
91 Santos JCS. Saturation of molecular genetic map in Urochloa humidicola (PhD Thesis). Campinas: University of Campinas; 2015. http://www.bibliotecadigital.unicamp.br/document/?code=000960129.
92 Fan C, Ma J, Guo Q, Li X, Wang H, Lu M. Selection of reference genes for quantitative real-Time PCR in bamboo (*Phyllostachys edulis*). Plos One. 2013;8:e56573.
93 Lambret-Frotté J, de Almeida LC, de Moura SM, Souza FL, Linhares FS, Alves-Ferreira M. Validating internal control genes for the accurate normalization of qPCR expression analysis of the novel model plant *setaria viridis*. PLoS One. 2015;10:e0135006.
94 Flügge U, Weber A, Fischer K, Lottspeich F, Eckerskorn C, Waegemann K, et al. The major chloroplast envelope polypeptide is the phosphate translocator and not the protein import receptor. Nature. 1991;353:364–7.
95 Fischer K, Arbinger B, Kammerer B, Busch C, Brink S, Wallmeier H, et al. Cloning and in vivo expression of functional triose phosphate/phosphate translocators from C3- and C4-plants: evidence for the putative participation of specific amino acid residues in the recognition of phosphoenolpyruvate. Plant J. 1994;5:215–26.
96 Koehler SM, Ho TH. Hormonal regulation, processing, and secretion of cysteine proteinases in barley aleurone layers. Plant Cell. 1990;2:769–83.
97 Mikkonen A, Porali I, Cercos M, Ho TH. A major cysteine proteinase, EPB, in germinating barley seeds: structure of two intronless genes and regulation of expression. Plant Mol Biol. 1996;31:239–54.
98 Furumoto T, Yamaguchi T, Ohshima-Ichie Y, Nakamura M, Tsuchida-Iwata Y, Shimamura M, et al. A plastidial sodium-dependent pyruvate transporter. Nature. 2011;476:472–5.

99 Belanger FC, Leustek T, Chu B, Kriz AL. Evidence for the thiamine biosynthetic pathway in higher-plant plastids and its developmental regulation. Plant Mol Biol. 1995;29:809–21.
100 Takahashi S, Yeo YS, Zhao Y, O'Maille PE, Greenhagen BT, Noel JP, et al. Functional characterization of premnaspirodiene oxygenase, a cytochrome P450 catalyzing regio- and stereo-specific hydroxylations of diverse sesquiterpene substrates. J Biol Chem. 2007;282:31744–54.
101 Bouché N, Scharlat A, Snedden W, Bouchez D, Fromm H. A novel family of calmodulin-binding transcription activators in multicellular organisms. J Biol Chem. 2002;277:21851–61.
102 Yang T, Poovaiah BW. A calmodulin-binding/CGCG box DNA-binding protein family involved in multiple signaling pathways in plants. J Biol Chem. 2002;277:45049–58.
103 Kim Y, Park S, Gilmour SJ, Thomashow MF. Roles of CAMTA transcription factors and salicylic acid in configuring the low-temperature transcriptome and freezing tolerance of *Arabidopsis*. Plant J. 2013;75:364–76. doi:10.1111/tpj.12205.
104 Hegarty MJ, Barker GL, Wilson ID, Abbott RJ, Edwards KJ, Hiscock SJ. Transcriptome shock after interspecific hybridization in senecio is ameliorated by genome duplication. Curr Biol. 2006;16:1652–9.
105 Josefsson C, Dilkes B, Comai L. Parent-dependent loss of gene silencing during interspecies hybridization. Curr Biol. 2006;16:1322–8.
106 Hirakawa H, Shirasawa K, Ohyama A, Fukuoka H, Aoki K, Rothan C. Genome-wide SNP genotyping to infer the effects on gene functions in tomato. DNA Res. 2013;20:221–33.