DNA Research, 2019, 0(0), 1–12
doi: 10.1093/dnares/dsz001
Full Paper

Full Paper

Unraveling the complex genome of *Saccharum spontaneum* using Polyploid Gene Assembler

Leandro Costa Nascimento[1,2], Karina Yanagui[1], Juliana Jose[1], Eduardo L. O. Camargo[1,3], Maria Carolina B. Grassi[1], Camila P. Cunha[4], José Antonio Bressiani[3], Guilherme M. A. Carvalho[5], Carlos Roberto Carvalho[5], Paula F. Prado[1], Piotr Mieczkowski[6], Gonçalo A. G. Pereira[1,*], and Marcelo F. Carazzolle[1]

[1]Laboratório de Genômica e bioEnergia (LGE), Departamento de Genética, Evolução, Microbiologia e Imunologia, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, SP, Brazil, [2]Laboratório Central de Tecnologias de Alto Desempenho (LaCTAD), Universidade Estadual de Campinas, Campinas, SP, Brazil, [3]Biocelere Agroindustrial Ltda, GranBio Investimentos S.A., Campinas, SP, Brazil, [4]Laboratório Nacional de Ciência e Tecnologia do Bioetanol (CTBE), Centro Nacional de Pesquisas em Energia e Materiais (CNPEM), Campinas, SP, Brazil, [5]Laboratório de Citogenética e Citometria, Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa, MG, Brazil, and [6]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

*To whom correspondence should be addressed. Tel. +55 19 3521 6237. Fax. +55 19 3521 6185.
Email: goncalo@unicamp.br

Edited by Prof. Kazuhiro Sato

Received 22 October 2018; Editorial decision 10 January 2019; Accepted 21 January 2019

Abstract

The Polyploid Gene Assembler (PGA), developed and tested in this study, represents a new strategy to perform gene-space assembly from complex genomes using low coverage DNA sequencing. The pipeline integrates reference-assisted loci and *de novo* assembly strategies to construct high-quality sequences focused on gene content. Pipeline validation was conducted with wheat (*Triticum aestivum*), a hexaploid species, using barley (*Hordeum vulgare*) as reference, that resulted in the identification of more than 90% of genes and several new genes. Moreover, PGA was used to assemble gene content in *Saccharum spontaneum* species, a parental lineage for hybrid sugarcane cultivars. *Saccharum spontaneum* gene sequence obtained was used to reference-guided transcriptome analysis of six different tissues. A total of 39,234 genes were identified, 60.4% clustered into known grass gene families. Thirty-seven gene families were expanded when compared with other grasses, three of them highlighted by the number of gene copies potentially involved in initial development and stress response. In addition, 3,108 promoters (many showing tissue specificity) were identified in this work. In summary, PGA can reconstruct high-quality gene sequences from polyploid genomes, as shown for wheat and *S. spontaneum* species, and it is more efficient than conventional genome assemblers using low coverage DNA sequencing.

Key words: sugarcane, genome assembly, transcriptome, gene discovery, new assembler

1. Introduction

Even with the progress of the next-generation DNA sequencing technologies, the *de novo* genome assembly remains a bottleneck. It is observed mainly in the case of complex genomes, such as plants, which are more difficult to assemble without a reference genome. Usually, plant genomes are larger than most mammals genomes,[1] have high polyploidy, contain a high number of repetitive regions and a high level of heterozygosity.[2] In addition, they contain a large number of gene families with numerous members and pseudogenes with nearly identical sequences due to effects of recent whole genome duplication and transposon activities. For these reasons, *de novo* assemblies of plants using only short reads generally create a highly fragmented genome split into hundred thousand scaffolds.[3] Although the cost of sequencing has declined exponentially in recent years made it feasible to achieve high sequencing coverage for plant genomes, this strategy has not shown enough to generate quality genomes, such as the cases of *Triticum aestivum*[4] and *Picea abies*,[5] which genome assemblies are highly fragmented with a sequencing coverage around 150×. Among the possibilities to increase the quality of these genomes, the development of new bioinformatics efforts can contribute to overcoming these constraints, especially, by exploring the continuous growth of closely related genomes and reference-assisted loci algorithms.[6]

Plants from *Saccharum* genus have very complex genome with high level of polyploidy, among them the commercial cultivars of sugarcane have an economic relevance due to high amount of sucrose. They are generated from several crossings between *Saccharum officinarum* and *Saccharum spontaneum*, followed by artificial selections of hybrids.[7] The genome profile of those hybrids consists of 70–80% of chromosomes from the *S. officinarum*, 10–20% of *S. spontaneum* and a few with inter-specific recombinations.[8,9] *Saccharum spontaneum* shows moderate sugar (8–10%) and high fibre (25–27%) contents, as well as high productivity (more than 180 t ha^{-1}) and resistance to biotic and abiotic stresses.[9,10] Moreover, *S. spontaneum* promises to improve fibre content in lignocellulosic feedstock that has an economically attractiveness for biofuels and other biochemicals production using recent technologies developed for biomass deconstruction.[10] The species is autopolyploid with chromosome numbers ranging from 40 to 128 and genome size ranging from 2.5 Gb to 12.5 Gb, with many aneuploidy forms.[8,11]

Due to its economic relevance of sugar and first-generation ethanol production, several groups have worked to disclose the molecular basis of commercial sugarcane varieties and significant contributions for genome and transcriptome profile were achieved. The SUCEST-FUN[12] and ORFeome[13] databases are great examples of large initiatives of researchers groups to understand the molecular biology of sugarcane. This database is a large platform that supports the development of System Biology projects of sugarcane providing information regarding the genome, gene expression and regulatory network. Apart from the SUCEST-FUN project, many other studies have made noteworthy contributions on the molecular basis of sugarcane.[14] Recently, a draft genome from sugarcane covering around 40% of the monoploid genome and more than 25,000 genes was published using a combination of BACs (Bacterial Artificial chromosomes), Illumina and PacBio sequencing.[15] However, most of these studies focused on commercial varieties of sugarcane, developed for the production of sugar and/or ethanol. A recent study of parental specie *S. spontaneum* investigated chromosomal structural rearrangements and disease resistance genes using whole genome sequencing.[16] However, the features related to productivity and accumulation of fibres were not analysed and are extremely necessary when we consider the production of biofuels from biomass and the transition for a green and sustainable economy.

In this work, we described a new and effective gene space assembly pipeline that allowed a reliable assembly of wheat (*T. aestivum*) and *S. spontaneum* genes using low coverage sequencing data. The identification of new genes, promoter regions, expanded gene families and transcriptome analysis of different tissues using reference-guided strategy were achieved. Our results showed the efficiency of Polyploid Gene Assembler (PGA) to reconstructed genes from complex genomes and contributed with new insights about the molecular basis of the *S. spontaneum* species, specifically involved with robustness and productivity of cane.

2. Material and methods

2.1. Plant material, genome and transcriptome sequencing

This study was performed using DNA and mRNA from *S. spontaneum* accessions US851008 and IN8482. Leaf, apical meristem, node, bud, internode and root tissues from three biological replicates (individual plants) were harvested from 9-month old and greenhouse-grown plants, immediately frozen in liquid nitrogen and stored at −80°C. For the DNA extraction, 100 mg of leaf tissue were grinded in liquid nitrogen using the Mini-Beadbeater-96 (Biospec Products, Bartlesville, EUA), for 1.5 min. Each sample was resuspended in 700 μl of Extraction Buffer (100 mM Tris-HCl pH8, 20 mM EDTA, 2% CTAB, 1% PVP 40, 1.4 M NaCl, 0.3% 2-mercaptoethanol) and incubated at 65°C for 30 min, homogenizing every 10 min. Then, 800 μl of chloroform–isoamylalcohol (24:1) was added to the tubes, gently mixed and centrifuged at 16,000 × *g* for 10 min. The supernatant (450 μl) was transferred to a new 1.5 ml tube, containing 120 μl of Extraction Solution (5% CTAB, 1.4 M NaCl), added with 570 μl of chloroform–isoamylalcohol (24:1). Samples were centrifuged at 16,000 × *g* for 10 min and the supernatant (400) transferred to a new 1.5 ml tube, followed by the addition of 400 μl isopropanol. Samples were centrifuged at 16,000 × *g* for 5 min and the pellet washed with 1 ml of 70% ethanol. Air-dried pellets were resuspended in 300 μl of sterile water. Each sample was incubated with 2 μl of ribonuclease (RNAse 10 mg ml^{-1}) at 37°C for 1 h. Quality control and quantification of the DNA were performed using dsDNA BR Assay (Invitrogen, Carlsbad, USA) and a Qubit fluorometer (Thermo Fisher Scientific, Waltham, USA), according to manufacturer's instructions. Total RNA (only from US851008) was extracted following the protocol described by Zeng and Yang (2002)[17] modified by Le Provost et al. (2007).[18] Quality control and quantification of RNA samples were performed using a Caliper LabChip XT micro capillary gel electrophoresis (Marshall Scientific, Hampton, USA) and a Qubit fluorometer (Thermo Fisher Scientific, Waltham, USA), respectively. Libraries were prepared with TruSeq Stranded mRNA Sample Preparation Kit (Illumina, San Diego, USA) using 1 μg of total RNA, according to manufacturer's instructions. Quality control of the libraries was performed as previously described. Samples were sequenced at the High-Throughput Sequencing Facility at the University of North Carolina (UNC, USA) using HiSeq2500 (Illumina, San Diego, USA).

2.2. Genome size estimation

Leaves from *Solanum lycopersicum* cultivar Stupicke, US851008 and IN8482 were processed for Flow Cytometry (FCM) at the

Laboratory of Cytogenetics and Cytometry (Federal University of Viçosa, Viçosa, Brazil). Genome size of *S. spontaneum* was measured from nucleic suspensions prepared as described by Carvalho et al. (2008).[19] The suspensions were analysed with a Partec PAS Cytometer (Partec_GmbH, Munster, Germany), equipped with a laser source (488 nm, for 2*C* value). *Solanum lycopersicum* was used as standard (2*C* = 2 pg). FCM parameters (e.g. gain and channel) were determined for each DNA sample based on external FCM analyses of primary standard and sample (data not shown). Each analysis was conducted with three technical replicates, accounting for more than 10,000 nuclei. 2*C*-value of *S. spontaneum* was calculated by dividing the mean channel of the G0/G1 fluorescence peak from the primary standard by that of the sample.

2.3. PGA pipeline

PGA was developed using PERL scripts for running in Linux system and integrates software to read mapping, *de novo* assembling and scaffolding. In mapping step, reads are split in sub-reads and mapped into the reference loci using Bowtie2[20] (v. 2.2.1). Assemblies are performed by Trinity software[21] (v. 2013-02-25) allowing contigs greater than 500 bp. In the scaffolding steps, L_RNA_scaffolder[22] (allowing introns up to 500,000 bp) and BLAT[23] (default parameters) are used for scaffolding using reference CDS and RNA sequences. SSPACE[24] (v. 2.0) with settings '–x 1-o 10 –v 2' is used for scaffolding using raw DNA-Seq paired-end reads.

2.4. PGA validation using *T. aestivum* genome

To validate PGA strategy, data from wheat (*T. aestivum*) were downloaded from Sequence Read Archive (SRA) from NCBI (National Center for Biotechnology Information) (Supplementary Table S1). The DNA-Seq reads were assembled with PGA using coverages of 3.6×, 5× and 7× (Supplementary Table S2) and the gene loci from barley (*Hordeum vulgare*) genome (24,243 sequences)[25] as reference. The wheat genes were downloaded from Phytozome[26] and compared with PGA assemblies using Exonerate[27] (v. 2.2.0) using the following settings '–refine region –refineboundary 100 –bestn 1 –percent 20 –quality 20 –geneseed 40 –seedrepeat 10 –minintron 30 –maxintron 500000'.

The gene prediction was performed using AUGUSTUS[28] (v. 3.2.1) with the following settings '–strand=both –genemodel=complete –species=wheat'. The AUGUSTUS results were compared with the wheat Phytozome gene prediction using BLASTn (*e*-value cutoff off 1e-20) and the 'No hits' sequences were selected. Among the 'No hits', we call new wheat genes those that met the following criteria: (i) non-aligned into the *T. aestivum* genome[4] with Exonerate[27] or, if aligned, lacking another predicted gene for at least 1,000 bp upstream or downstream by the windowBed tool from BEDTools[29] (v. 2.17.0); (ii) have a hit in the Uniref90 database,[30] using BLASTp (*e*-value cutoff off 1e-5).

2.5. *Saccharum spontaneum* gene space assembly

Saccharum spontaneum DNA-Seq reads (Supplementary Table S3) were assembled by PGA using *Sorghum bicolor*,[31] *Zea mays*[32] and *Setaria italica*[33] genes as reference, chosen based on the phylogenetic distance. The gene space assembly was here referred as '*Saccharum spontaneum 1.0*'. To comparison, *S. spontaneum* reads were assembled by traditional approach using SOAPdenovo,[34] configured to use 73 K-mer, chosen by N50 maximization.

2.6. Splicing alignment of *S. spontaneum* RNA-Seq reads

Paired-end and single-end reads from 18 RNA-Seq libraries (leaf, root, internode, node bud and apical meristem) were mapped against '*Saccharum spontaneum 1.0*' using TopHat[35] (v. 2.0.9) configured with the settings '–i 10 –I 500000 –library type fr-firststrand'. The percentage of mapped reads on each sample is available at Supplementary Table S4. RNA-Seq data from wheat and maize were also mapped against their own reference genomes using software and parameters described above (Supplementary Table S5).

2.7. *De novo* transcriptome assembly and annotation

Trinity[25] (v. 2013-02-25) was used to perform *de novo* transcriptome assembly using 18 RNA-Seq libraries using the settings '–seqType fq –JM 100G –normalize_reads –normalize_max_read_cov 30'. Larger transcripts from each Trinity component were annotated using BLASTx (*e*-value cutoff >1e-5) against the Uniref9030 and Swiss-Prot[36] databases (Supplementary Table S6).

2.8. *Saccharum spontaneum* gene prediction

Gene prediction in '*Saccharum spontaneum 1.0*' was refined using the following datasets as extrinsic evidence: (i) 215,681 proteins from plants: 33,012 from great millet (*S. bicolor*),[31] 80,713 from maize (*Z. mays*),[32] 35,471 from foxtail millet (*S. italica*),[33] 39,049 from rice (*Oryza sativa*)[37] and 27,436 from *Arabidopsis thaliana*[38]; (ii) 626,769 transcripts from *Saccharum* spp.: 358,695 from *S. spontaneum* (Section 4.7) and 268,034 from sugarcane[13,39] and (iii) splicing alignments (BAM files) (Section 4.6). Datasets (i) and (ii) were mapped to *S. spontaneum* genome assembly using Exonerate[27] (v. 2.2.0) with the following settings '–refine region –refineboundary 100 –bestn 1 –percent 40 –quality 40 –geneseed 40 –seedrepeat 10 –minintron 30 –maxintron 500000'. The Exonerate[27] alignments were converted to GFF format with a custom PERL script. The BAM files from dataset (iii) were processed by bam2hints tools, from AUGUSTUS[28] package (v. 3.0.1) with the option '–intronsonly', to convert the BAM files to GFF format. A custom PERL script filtered the hints files to retain only evidences of introns supported by three or more read alignments.

The prediction of the coding loci was also performed using AUGUSTUS[28] trained and configured as '–strand=both –genemodel=complete –alternatives-from-evidence=true –gff3 = on'. The training set was formed by 1,000 transcripts fully covered by one protein (the transcript must contains START and STOP codons in the correct position identified through comparison of BLASTx against the Uniref90 database[30]) from the Trinity[21] *de novo* assembly (Section 4.7). GFF files generated by Exonerate[27] and bam2hints programs were used as evidences of exons and introns. Predicted genes were mapped against CDD (Conserved Domain Database) using RPSBLAST and proteins with similarity superior to 30% to transposons (TE) domains were removed our dataset. The filtered gene models were submitted to PASA pipeline[40] (v. 2.0.2) to improve the identification of untranslated regions (UTRs) and new splice variants.

The confidence of the predicted genes was determined by comparing coding sequences to 626,769 transcripts from *Saccharum* genus using BLASTn (*e*-value cutoff <1e-5), with transcripts classified according to the alignment coverage: high confidence (HC; gene alignment coverage ≥60%). Medium confidence (MC; 30% ≤ alignment coverage ≥60%) and low confidence (LC; < 30% alignment coverage or no hit). Finally, BUSCO pipeline[41] was performed using

the plant dataset (Embryophyta odb9) to verify full-length HC proteins and compare their sequences with other homeologues in complex genomes, maize and sugarcane.

2.9. Read depth coverage

To identify coverage from the gene regions in '*Saccharum spontaneum 1.0*', the raw DNA reads (Supplementary Table S3) were compared with the *Saccharum spontaneum 1.0* assembly using Bowtie2[20] (v. 2.2.1) using default parameters. The tools coverage and genomecov from BEDTools[29] (v. 2.17.0) were used to extract the read depth coverage from the alignment results (BAM files).

2.10. Functional annotation of predicted genes

The functional annotation of predicted *S. spontaneum* proteins was performed using BLASTp (*e*-value cutoff <1e-5) against several databases, including Swiss-Prot,[36] Uniref9030, NR from NCBI, TAIR[38] and CDD. Gene ontology terms association and enrichment analysis were performed using BLAST2GO pipeline.[42]

2.11. Comparative genomics

We used the proteins from assembled genome sequences of five Poaceae species available from the Phytozome v.12 database[26]: maize (PH207 v1.1),[32] great millet (v3.1.1),[31] foxtail millet (v2.2),[33] rice (v7)[37] and purple false brome (*Brachypodium distachyon* v3.1)[43]; and bamboo (*Phyllostachys heterocycla*) (EMBL, accession ERP001340).[44] Homeologue gene groups were assigned between those species and *S. spontaneum* using the Markov clustering algorithm implemented in OrthoMCL.[45] Single-copy orthologue (SCO) genes were used to a phylogenetic reconstruction. SCO genes were aligned individually with multiple alignment algorithms implemented in MAFFT[46] using the iterative refinement method and WSP and consistency scores (G-INS-i). All alignments were concatenated in a supermatrix for the maximum likelihood phylogenetic inference in RAxML v8[47] with the GTR+GAMMA model of substitutions and 1,000 bootstrap replicates for branch support. Gene groups with size changes during *S. spontaneum* evolution were identified by BadiRate[48] using the phylogenetic history, providing evolutionary inferences for expansion and retraction of gene families.

2.12. Transcript expression calculation

The BAM files from RNA-Seq alignment to '*Saccharum spontaneum 1.0*' were inputted into Cufflinks[49] (v. 2.1.1) to produce transcript assemblies, one for each sample. We used the settings '–I 500000 –multi-read-correct' and the GFF file of *S. spontaneum* gene prediction to guide the assemblies. Subsequently, the cuffmerge/cuffcompare tools merged 18 cufflinks assemblies with the settings '–min-isoform-fraction 0.1'. The cuffcompare tool was used to classify the genes into new *S. spontaneum* isoforms, a transcript in the opposite strand than reference gene and as an intergenic transcript. The raw RNA-Seq reads were aligned against the transcriptome assembly using RSEM[50] (v. 1.2.19) with the parameter '–strand-specific' to perform multi-mapping alignments and estimate the expression values for each transcript using RPKM (Reads Per Kilobase per Million mapped reads).

2.13. ncRNA identification

ncRNAs were identified in the transcriptome assembly (Section 4.12) assuming four criteria: (i) classified as intergenic or in opposite strand than other transcript; (ii) no hits against NR database from NCBI (BLASTx *e*-value cutoff <1e-5); (iii) Coding Potential Calculator (CPC)[51] score ≤ -1 (although CPC classifies ncRNA as transcripts with values ≤ 0, we decided to be more conservative to avoid false positives) and (iv) genomic location larger than 1,000 bp from scaffold tips to avoid partial genes not previously predicted.

2.14. Promoter sequences analysis

A PERL script was developed to extract the start position of the alignment (first hit) from BLASTp/Uniref9030 output and, then, evaluate whether the predicted genes would be equivalent to complete protein prediction from the database. Because of the low conservation of the signal peptide, a margin of 20 amino acids was allowed at the beginning of each alignment. If the alignment started at up to 20 amino acids after the start codon, the protein beginning was in the correct position. The complete proteins were searched for the presence of at least 2,000 bp upstream of the start codon and those identified were subjected to a tissue-specific analysis using TAU metric.[52] TAU scores above 0.95 and RPKM average above 10 were potential candidates for tissue-specific promoters. A list with constitutive promoters was generated by calculating the coefficient of variance (standard deviation/mean) of each gene and it was considered constitutive when that coefficient was below 15%. Standard deviation and mean for each gene was calculated over RPKM values among all tissues.

2.15. *Saccharum* spp. databases comparison

HC genes from *S. spontaneum* identified in this study were compared with three datasets of sugarcane genes: gene prediction from the sugarcane genome[15] (25,316 sequences), SUCEST-FUN[12] (43,141 sequences) and ORFeome[13] (195,765 sequences) databases using BLASTn (*e*-value cut-off $\leq 1e^{-5}$). 'No hits' sequences were selected by a custom PERL script and the annotation from Uniref9030 from each one was selected. The expression profile was subjected to a clustering analysis using Seaborn's clustermap package from Python.

3. Results and discussion

3.1. Polyploid Gene Assembler

PGA pipeline works with whole shotgun sequencing data of polyploidy and other complex genomes and focuses on gene assembly, including exons, introns, UTRs and promoters. It performs two main steps: (i) a reference-assisted assembly (here refereed as Reference Loci Assembly) and (ii) a *de novo* assembly with non-used reads in the first step. The pipeline integrates various software for read mapping, *de novo* assembling and scaffolding as described in Fig. 1A. PGA can be executed on a desktop machine with RAM memory configured according to the computational power available. Better performance is achieved using a server with at least 16 CPU cores, 64 GB of RAM memory and 500 GB of Hard Disk. Updates, new versions and results from PGA are available at the web site: http://www.lge.ibi.unicamp.br/pga (date last accessed 25 January 2019). PGA pipeline is valuable for any polyploid organism that has phylogenetically close-related species with available complete genome sequences.

PGA Reference Loci Assembly step. DNA-Seq reads are divided into three overlapping pieces with sizes equal to 50% of the original read size, called sub-reads (Supplementary Fig. S1). The overlap between sub-reads prevents reads with tip differences (mismatches or gaps) from being unmapped, improving alignment rate in low conserved regions or intron–exon junctions from reference. The sub-

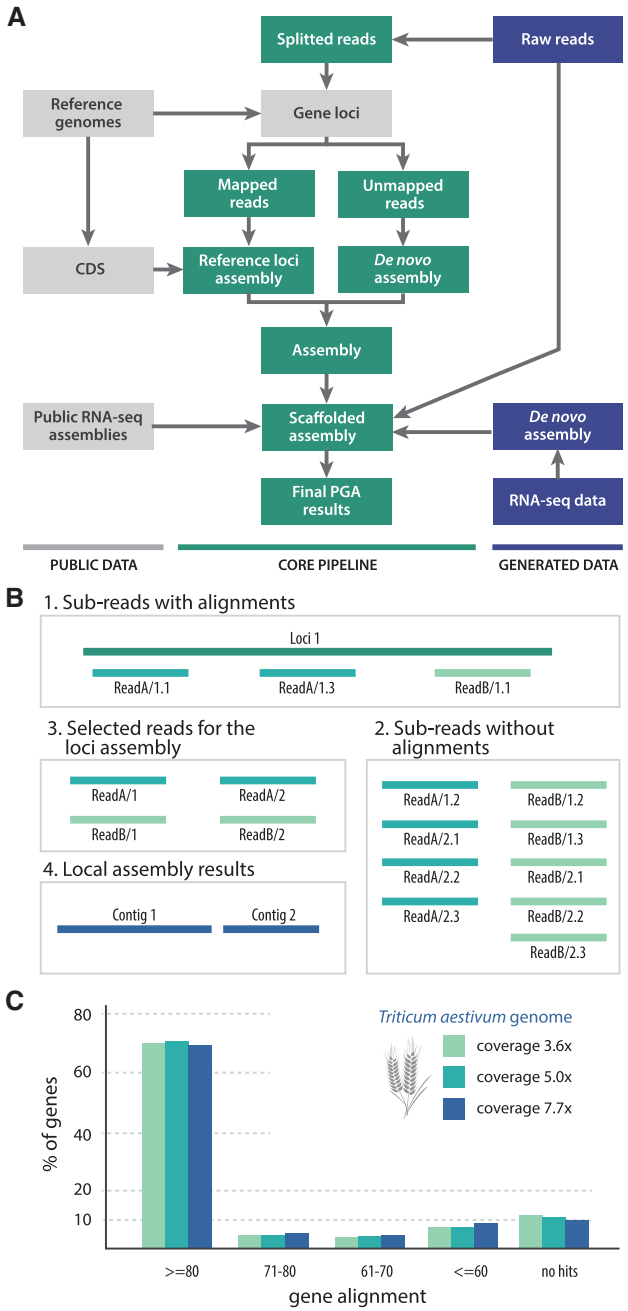


Figure 1. The PGA pipeline. (A) The three steps of the PGA pipeline are outlined, summarizing the methodology developed in this work. (B) The Reference Loci Assembly step of one locus (Loci 1) is show. Only three sub-reads (ReadA/1.1, ReadA/1.3 and ReadB/1.1) from two read-pairs (A and B) were mapped into the sequence of the Loci 1 (1 and 2), but the pipeline uses the both pairs (whole reads) in the local assembly (3) producing two contigs (4). (C) Validation of the PGA using sequencing data of *T. aestivum*.

reads are mapped against a selected reference loci, usually an evolutionary close-related species with available genome data. Reference loci must include exons, introns, UTRs and up to 500 bp upstream and downstream UTRs from one or more species to gene discovery maximization. On this step, PGA can be configured to allow differences (default is up to 10% of gaps or mismatches) between the sub-read and the reference sequence. Full reads corresponding to sub-reads mapped to a locus are retrieved and included in a *de novo* assembly performed to each locus separately (Fig. 1B). In cases of sub-reads mapping to different loci, the locus with the highest number of occurrences is selected. Original CDS sequences from the reference(s) genome(s) can be used for scaffolding. This step is optional; however, a significant increase in longer contigs is observed (contigs ≥5,000 bp) (Supplementary Table S7).

PGA de novo Loci Assembly and Scaffolding steps. *De novo* assembly is performed only with reads that were discarded in the first step. Due the complexity of the polyploid genomes, this step generates very fragmented contigs (Supplementary Table S8). Despite that, we consider *de novo* assembly very important to identify genes that are exclusive in the studied organism. Optionally, is possible to perform a new round of scaffolding using raw DNA-Seq reads applied to *de novo* assembled contigs and reference-loci assembled contigs. After scaffolding step, it is possible to use RNA assemblies (assembled RNA-Seq or ESTs) to merge contigs. This step is also optional, depending on the availability of transcriptomic data of the organism of interest.

PGA uses a transcriptome assembler (Trinity[21]) instead of traditional genome assembler due to its best performance in comparison with conventional genome assemblers in most of the cases tested (Supplementary Table S9). It happened, probably, because Trinity was developed to support variations of coverage over the locus once it assembles the transcripts generated by alternative splicing events. All traditional genome assemblers, such as SOAPdenovo[34] and Velvet,[53] are based on concept of uniform coverage, varying only on repetitive regions, and stop the formation of a contig when coverage changes significantly. This concept cannot be applied to polyploid genome, because homeologous chromosomes have regions with low (intergenic) and high (mainly exons) similarities,[10] which causes abrupt changes in coverage (Supplementary Fig. S2).

The PGA strategy was validated using public available DNA and RNA sequence data (SRA, NCBI, Supplementary Table S1) from *T. aestivum* subsp. *aestivum* (bread wheat), a hexaploid species (size of 17 Gb) with high amount of near identical or duplicated sequences[4] using three different genome coverages (3.6, 5.0 and 7.7×) (Supplementary Table S2) and *H. vulgare* (barley) genome as reference 25. A total of 99,386 wheat genes available at Phytozome[26] were used for comparison with the PGA scaffolds and resulted in almost 70% identified with up to 80% of alignment coverage (Fig. 1C). In addition, the most part of the partially aligned genes (<80% of alignment) have a very similar sequence but truncated in the borders (Supplementary Fig. S3), i.e. our assembly failed in the beginning or the end of the loci, possibly due the short scaffold sequences. Finally, the PGA assembly with coverage of 3.6× was submitted to gene prediction using AUGUSTUS[28] to verify if the pipeline was able to identify new wheat genes. This analysis resulted in a total of 6,378 predicted genes in the PGA assembly that were not identified by the Phytozome[26] prediction. From these, 2,032 are probably new wheat genes because match in two criteria: (i) all of them have a BLASTp plant hits against Uniref9030 database; (ii) 171 genes did not align into the *T. aestivum* genome (Supplementary Fig. S4) and 1,861 genes aligned into genome, but in the intergenic regions, lacking another predicted gene for at least 1,000 bp upstream or downstream (Supplementary Fig. S5). All these genes are new contributions for the complex wheat genome obtained by PGA pipeline using a very low sequencing coverage. The remaining 4,346 could be also real genes, but were discarded because did not match to one or the both criteria used.

Table 1. Summary of three steps of PGA pipeline applied to *S. spontaneum*, including the *Saccharum spontaneum 1.0* assembly

	Reference assembly	*De novo* assembly	After scaffolding (*Saccharum spontaneum 1.0*)
Sequences ≥ 500 bp	81,998	132,518	179,429
Max sequence length (bp)	30,266	26,551	53,214
Average sequence length (bp)	1,720	1,053	1,597
N50 (bp)	2,789	1,130	2,223
Sequences ≥ 1,000 bp	39,213	49,390	79,690
Sequences ≥ 5,000 bp	4,921	454	9,098

3.2. Genome sequencing and gene-space assembly of *S. spontaneum*

There is a large variation in genome size among the varieties of *S. spontaneum* species ($2n$ = 40–128).[11] Therefore, the genome size of two *S. spontaneum* accessions, US851008 and IN8482, was experimentally estimated by flux cytometry resulting in 9.3 and 11.8 Gb, respectively (Supplementary Fig. S6). Based on this result, DNA libraries of the smallest genome, i.e. *S. spontaneum* accession US851008, were sequenced by Illumina technology that produced 60 Gb of 100-bp paired end reads (Supplementary Table S3), representing a low coverage DNA sequencing of 6.4×. In addition, RNA-Seq libraries from leaf, apical meristem, node, bud, internode and root tissues were sequenced and produced around 20 million of 100-bp paired-end reads and 50 million of 100-bp single-end reads per library (Supplementary Table S4).

The application of PGA to *S. spontaneum* reads was performed using three closely related genomes as reference: *S. bicolor*,[31] *Z. mays*[32] and *S. italica*.[33] Based on the phylogenetic distance, the 33,012 *S. bicolor* loci were used first, followed by 80,713 *Z. mays* loci and finished with 35,471 *S. italica* loci. The pipeline was executed for 23 days and used a maximum memory of 60 GB. PGA produced 79,690 scaffolds >1,000 bp and this assembly was here referred as '*Saccharum spontaneum 1.0*'. Despite the '*Saccharum spontaneum 1.0*' represents a high-fragmented draft genome, as observed by the number of sequences (≥500 bp) after scaffolding (179,429) and average sequence length (1,597 bp) (Table 1), it was expected that most of the sequences represent different protein-coding genes. To reinforce this hypothesis, the RNA-Seq reads were mapped against '*Saccharum spontaneum 1.0*' resulting in around 70% of matches (Supplementary Table S4), which it is similar value in comparison with wheat (average of 74.3%) and a close value in comparison with maize (average of 84%) (Supplementary Table S5).

3.3. Gene prediction and annotation

A total of 53,436 genes and 59,890 transcripts were predicted on '*Saccharum spontaneum* 1.0' using a combination of software and external evidence using proteins from closely related organisms and transcripts from *Saccharum* genus (as described in Section 2). The predicted proteins were annotated against public protein databases, resulting in more than 80% and 50% of genes supported by protein hit on the Uniref9030 and Swiss-Prot,[36] respectively (Fig. 2A). Also, the predictions were compared with grasses genes indicating high similarity with *S. bicolor*[31] and *S. italica*[33] (Supplementary Fig. S7). The 53,436 predicted genes in *Saccharum spontaneum 1.0* were classified considering the alignment coverage with 626,769 transcriptome sequences, resulting in 39,234 HC genes (alignment coverage ≥60%), 7,322 MC genes (30% ≤ alignment coverage ≥60%) and

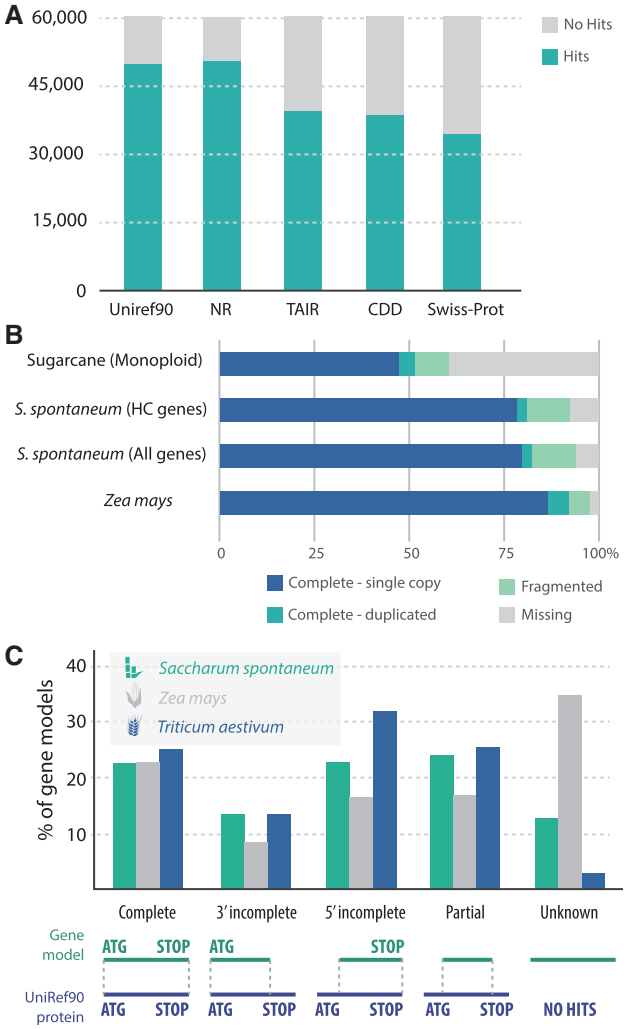


Figure 2. *Saccharum spontaneum* gene prediction. (A) Annotation results from all genes against public protein databases. (B) BUSCO results using genes from sugarcane, *S. spontaneum* and *Z. mays*. (C) Accessing the completeness of the gene prediction from *S. spontaneum*, *Z. mays* and *T. aestivum*.

6,880 as LC genes (2,903 genes with an <30% alignment coverage and 3,977 no hits). A summary of the HC genes is shown in Table 2.

To evaluate the completeness of HC proteins, the BUSCO[41] analysis was applied using a set of 1,440 highly conserved plant orthologous proteins. The analysis included *Z. mays*[32] and Sugarcane proteins.[15] A majority (90%) of the 1,440 proteins in the plant BUSCO database was covered by HC proteins, and of these

Table 2. Summary of *S. spontaneum* gene prediction (only HC genes are shown)

#Gene loci	39,234
#Transcripts	43,462
Mean CDS length (bp)	902.85
Max./min. CDS length (bp)	14,859/123
Mean transcript length (bp)	1,082
Mean exon length (bp)	247
Max./min. exon length (bp)	6,702/3
Single exon genes	9,683
Swiss-Prot database support (>30%/60%)	23,347/20,121
Uniref90 database support (>30%/60%)	33,945/32,148

80% were classified as complete and single copy. This result was similar to the *Z. mays* and better than Sugarcane proteins (Fig. 2B). Once the BUSCO analysis is based on a small set of single copy orthologous proteins, the completeness of HC proteins was expanded to all annotated proteins (BLASTp against Uniref9030) that were classified based on the alignment coverage: (i) 'Complete' predicted gene, when the predicted protein (query) covers very well the database subject (first hit from database); (ii) '3′ incomplete' genes, when the predicted protein covers only the begin of database subject; (iii) '5′ incomplete' genes, when the predicted protein covers only the final of the database subject and (iv) 'Fragmented' genes, when there is partial alignment on both sides. For comparison, the same procedure was performed using predicted proteins from *Z. mays*[32] and *T. aestivum*[4] genomes available at Phytozome,[26] chosen for the complex genomes in the grass family. A larger amount of complete predicted genes in *S. spontaneum* were observed when compared with *Z. mays*, a largely studied genome, and *T. aestivum* (Fig. 2C).

PGA can assembly gene regions in polyploid organism using low sequencing coverage, because the high level of sequence conservation among homeologous genes increase the sequencing coverage in these regions. To show this effect, we calculated the sequencing coverage over gene regions by alignment of *S. spontaneum* reads against '*Saccharum spontaneum 1.0*' and counting the read depth for each base pair in that regions (see Section 2). As shown in Supplementary Fig. S8, the distributions of read depth by base pair in the gene regions considering all predicted genes have peak at 82×. This new sequencing coverage represents ∼12 times the value previously calculated (6.4×) using the whole genome size and, interestingly, near the expected value of ploidy for *S. spontaneum*.

For comparison purposes, the *S. spontaneum* reads were assembled by conventional strategy using SOAPdenovo[34] (Table 3). The SOAPdenovo was executed 12 days and used a maximum memory of 448 GB, 7.5× more memory than PGA. The assembled sequences were submitted for gene prediction using AUGUSTUS[28] (trained for *S. spontaneum* as described previously) that identified 54,658 genes. Despite this number is very similar than that obtained using PGA, the mean size of them (564.1 bp) was smaller than our dataset (53,436 genes with mean size of 840.3 bp). In addition, we performed completeness analysis using BUSCO[41] on SOAPdenovo gene prediction that showed a small covered over the BUSCO dataset (34.2%), being the majority fragmented (21.4%). These results reinforce the strategy based on reference-assisted loci that was used.

In summary, the PGA pipeline produced high-quality *S. spontaneum* sequences focused on gene content in comparison with conventional genome assemblers (Table 3) and with other published complex genomes (Fig. 2). The high-coverage sequencing over the gene regions, the use of transcriptome assembler and the criteria chosen for sequence selection (see Section 2) generated a representative sequence for each locus, avoiding erroneous duplication of homeologous alleles, that can be observed by low values of duplicated proteins (2.8%) in BUSCO[41] analysis.

Table 3. Comparison between *S. spontaneum* assemblies using PGA and SOAPdenovo

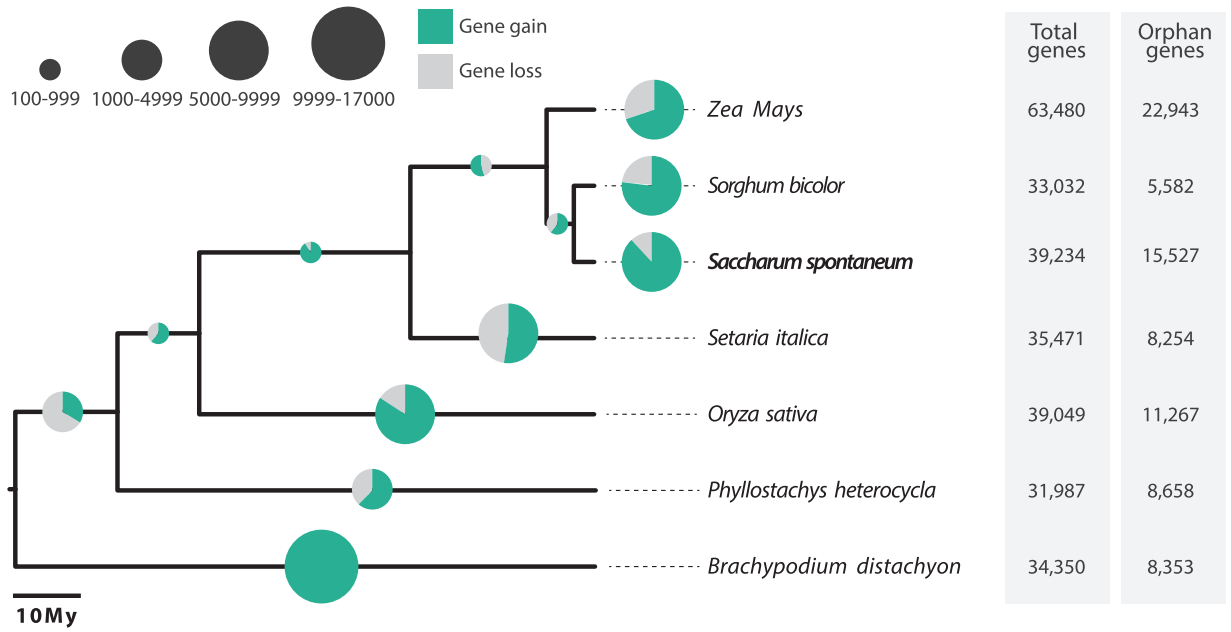
	PGA	SOAPdenovo
Scaffolds (≥ 1,000 bp)	79,690	135,024
Mean scaffold size (bp)	2,720	1,733
N50 (bp)	3,491	1,744
Largest scaffold (bp)	53,214	54,032
Total size (bp)	216,774,882	234,014,899
RAM memory used	60 GB	448 GB
Time to run	23 days	12 days
Gene loci	53,436	54,658
Mean size of genes (bp)	840.3	564.1
Complete and single copy (BUSCO)	79.6%	11.5%
Complete and duplicated (BUSCO)	2.8%	1.3%
Fragmented (BUSCO)	11.5%	21.4%
Missing (BUSCO)	6.1%	65.8%

3.4. Comparative genome and gene family analysis

From the total of 39,234 HC genes predicted for *S. spontaneum*, 23,707 was clustered into gene families with genes from at least one other species and 15,527 genes were not assigned to any family and, thus, considered orphans (Fig. 3). SCOs appeared in 5,500 clusters that were used to infer the phylogeny in Fig. 3.

Saccharum spontaneum showed a high amount of orphan genes, like *Z. mays* e *O. sativa*, and 80% of them were assembled in the *de novo* step of our pipeline showing the importance of this step to reconstruct species-specific genes. Orphan genes represent genes that are in either a neutral process of pseudogenization or an adaptive process of evolutionary novelties of the lineage. Many orphan genes in plants have been reported as evolutionary novelties related to environmental stress response and species-specific traits[54] and the large amount of orphan genes in *S. spontaneum* may be also related to traits selected during domestication. We evaluated the gene expression profile of the orphan genes and 52.5% have RPKM >= 3 in at least one RNA-Seq libraries, being 31.8% with highly expression values (RPKM >= 10). In addition, the most of 15,527 genes (53.3%) have similarity with the gene prediction from sugarcane (BLASTn, *e*-value cutoff of 1e-10) with some of them covering more than 90% of the sugarcane gene, such as: SS46658 (annotated as Peroxidase), SS21041 (annotated as Alpha-galactosidase) and SS4010 (annotated as transcriptional factor).

A total of 29,268 gene families with two or more genes were obtained, revealing 14,965 gene families with recent paralogues in at least one species, 138 families with recent paralogues exclusive to *S. spontaneum*, summing up 143 new duplication events in *S. spontaneum*: 133 duplications and 5 triplications (Supplementary Table S10). We also found 238 exclusive or expanded (or exclusive/-expanded) gene families from *S. spontaneum*, with ∼95% representing unknown-domain protein or no hits in blast searches against NR/NCBI database (Supplementary Table S11). We detected expression in 122 out of 167 genes among no hits families in at least one RNA-Seq library, which indicates that those families are not artefacts of the technique.



	Total genes	Orphan genes
Zea Mays	63,480	22,943
Sorghum bicolor	33,032	5,582
Saccharum spontaneum	39,234	15,527
Setaria italica	35,471	8,254
Oryza sativa	39,049	11,267
Phyllostachys heterocycla	31,987	8,658
Brachypodium distachyon	34,350	8,353

Figure 3. Phylogenetic inference of the relationships among Poaceae species was compared with *S. spontaneum*. Phylogeny was obtained by Maximum Likelihood analysis using a concatenated matrix of the 5,500 single-copy gene families obtained in OrthoMCL. All branches got total support (100%) for the 1,000 bootstraps used. The total number of genes and the orphan genes for each species is on the right-side table. For each branch in phylogeny a pie chart reflects the proportion of gene gains and losses estimated in BadiRate analysis. The size of the pie charts reflects the absolute number of genes gained or lost for all families in that current or ancestral lineage.

Expansions and retractions of gene families, analysed using phylogeny and the birth and death models, were used to calculate rates of gain and loss of genes through evolution (Fig. 3 and Supplementary Table S11). In general, Poaceae species showed a higher gene gain than loss, with absolute numbers varying from 5,000 to 10,000 genes for each lineage. All ancestral lineages showed a lower amount of both gain and loss, suggesting that the current lineages have higher gene innovations. Those patterns of innovation may be associated with the domestication history of each crop. The highest proportion of gains was found within the grass clade, grouping *S. spontaneum*, *S. bicolor*, *Z. mays* and *S. italica*.

A total of 37 gene families were expanded in *S. spontaneum* when compared with its ancestral lineages. Most of them (30) were classified as no hits, unknown function, mobile element or ribosomal related in blast searches against NR/NCBI databases, although expression evidence was found in the RNA-Seq data. Three families were selected as promising candidates of expansion in *S. spontaneum* due to the number of genes, sequence completeness and agronomical relevance: leucine-rich repeat receptor-like kinase (LRR-RLK) family (ORTHOMCL0, 35 putative evolutionary novelties out of 61 genes), GRAS-domain protein family (ORTHOMCL18, 9 out of 26) and callose synthase (CalS) family (ORTHOMCL23, 12 out of 19) (Supplementary Figs S9 and S10; Table S12).

The LRR-RLK and GRAS domain are multigene and complex gene families, regulating many features in plant growth and development and stress response. LRR-RLK family encodes cell surface receptors responsible for cell–cell and cell–environment communication, affecting axillary and shoot apical meristem (SAM), tillering and brassinosteroid signalling pathway.[55] Whereas, GRAS domain protein family modulates shoot and root development, through gibberellin and light signal transduction, axillary meristem initiation, SAM maintenance and root radial patterning.[56] Recently, ScGAI, a GRAS protein and hormonal hub from sugarcane, was found in SAM and elongating internodes, implicated in shoot-root ratio modulation and tillering.[57] Although difficult to devise gene function through phylogeny, the LRR-RLK and GRAS domain gene novelties from *S. spontaneum* showed on average higher gene expression levels in node and/or buds (axillary meristem) (Supplementary Fig. S9), indicating a potential role in initial growth development, coherently with the species capacity for intense tillering and faster culm development.

Callose, a minor cell wall polymer, is usually associated with specialized tissues (cell plate during cytokinesis, vascular system, pollen and pollen tube formation and plasmodesmata) and response to pathogen and herbivore attacks.[58] Although composed mainly of cellulose, callose is found in significant amounts in cell walls of energy crops, such as maize and *Miscanthus x giganteus* (up to 5%) leaves, in which callose fibrils intertwined in cellulose forms an atypical outer layer.[58] Callose-enriched biomass is a new target for biotechnological engineering resulting in significant increase in second generation ethanol yield using an optimized system of enzymes and yeast.[59] In this study, putative CalS gene novelties were ubiquitously and highly expressed (RPKM values from 103 to 174; Supplementary Fig. S11) in all tissues evaluated and might represent new and interesting targets for sugarcane and energy cane breeding.

3.5. Reference transcriptome assembly

The transcriptome assembly was produced using the Cufflinks[49] package using *Saccharum spontaneum 1.0* as the reference. As a result, 156,531 transcripts were identified with a length >200 bp. The comparison between the reference and *de novo* transcriptome assembly, shown in Supplementary Table S13, reveals that a set of higher quality sequences is constructed when the '*Saccharum spontaneum 1.0*' is used to guide the transcriptome assembly. The transcriptome assembly using our reference genome proved to return high-quality and more informative data, with longer transcripts and ORFs, and

more hits against Swiss-Prot.[36] In addition, there is high concordance in gene structure, such as exon–intron junctions and intron length, between gene predictions and splicing alignments of RNA-Seq reads, as shown in Supplementary Fig. S12.

In addition to improving the quality of the transcripts, the use of *Saccharum spontaneum 1.0* allows the identification of candidates for isoforms and long non-coding RNAs. For that the transcripts identified as a transcript in the opposite strand than reference gene (5,566) and as an intergenic (41,931) by the cuffcompare were submitted to the ncRNA identification. A total of 3,178 transcripts (mean size: 624 bp) could be assigned as non-coding (2,822 bigger than 200 bp, i.e. lncRNAs) (Supplementary Table S14). As expected, those transcripts have little expression values, as shown in Supplementary Fig. S13, in all RNA-Seq samples around 50% of non-coding RNAs have RPKM > 1 and around 20% have RPKM > 3.

3.6. Promoter sequences

PGA pipeline can extend the scaffold sequences upstream from the gene, providing information on the promoter regions that can be used for many biotechnology applications. The use of constitutive or tissue-specific promoters is essential to carry out studies to investigate plant metabolism under defined conditions. Moreover, the use of right promoter to express heterologous proteins in a controlled condition and specific tissues is mandatory for the development of genetically modified varieties.

In this study, a total of 3,108 promoter regions were identified considering at least 2,000 bp upstream from the beginning of the gene. The initial positions of predicted genes were verified by alignment against Uniref90 database (see Section 2). These promoter regions were classified according to their respective gene expression profiles (constitutive or tissue-specific) and intensities: weak (10 < RPKM < 100) and strong (RPKM >= 100). A summary of the promoter regions identified is shown in Table 4 and reveals that the most represented group is formed by constitutive promoters (Supplementary File S1) and there are many interesting tissue-specific promoters distributed among leaf (Supplementary File S2), root (Supplementary File S3), node (Supplementary File S4), bud (Supplementary File S5) and apical meristem (Supplementary File S6).

3.7. Comparison with sugarcane databases

To identify *S. spontaneum* genes that were not described previously, we compared the HC genes produced by *Saccharum spontaneum 1.0* with the genes from the Sugarcane genome[15] (25,316 sequences) and the transcriptome sequences from SUCEST-FUN[12] (43,141 sequences) and ORFeome[13] (195,765 sequences) (Fig. 4A), resulting in 1,913 genes that were identified only in our assembly (Supplementary Table S15). From those genes, 536 have known functions with proteins already characterized in other plants (mainly monocots), 502 were classified as hypothetical proteins, and 926 were no hits (Fig. 4A). Moreover, the expression values of the 1,913 genes in *S. spontaneum* (Fig. 4B), show distinct pattern across the selected tissues (apical meristem, lateral bud, internode, node, leaf and root).

Saccharum spontaneum is one of the most robust grasses, having a high tolerance to abiotic and biotic stresses, mainly drought resistance and the ability to grow in nutrient-poor soils.[9] Remarkable, from the genes presenting high expression values (≥50 FPKM), several are involved in stress response and 135 new genes were identified in roots. The ability to cope with several stresses of *S. spontaneum* could be a consequence of its large and dense root system when compared with *S. officinarum* and the commercial varieties of sugarcane, which make possible the soil exploitation and the higher uptake of water and nutrients. In addition, several genes could be related to symbiosis establishment (Table 5), recently described to increased drought resistance of *S. spontaneum*.[66] Among those genes, we can highlight genes related to recognition and early steps of symbiosis establishment (SS34074, SS37878, SS50399); stress and disease response (SS11486, SS38841, SS35410, SS41259) and cell wall degradation (SS48435); electron transfer and primary metabolism (SS35024, SS38176). Genes coding for nutrient transporter between fungus and plant were also identified (SS28265, SS11782, SS31510, SS13699), although they were already annotated in SUCEST-FUN[12] and ORFeome[13] databases.

Between other new genes, expressed exclusively in roots cells, we identified one corresponding to a Late Embryogenesis Abundant (LEA) (SS1022; 28.01 ± 4.12 RPKM). Characterized as a hydrophilic protein widely distributed in the plant kingdom, LEA is accumulated in vegetative tissues as a response to water limitation due to salinity, drought or high temperatures.[67] Other significant genes annotated by our analysis and expressed only in *S. spontaneum* roots are the BURP domain-containing genes, specifically; those related to families 10 and 13 (BURP10 and BURP13). BURP domain-containing genes are related to many functions in plants, in *O. sativa*, the BURP10 was reposted to be highly expressed in roots at the tillering stage.[68] Our results also demonstrated the expression of BURP10 (SS23435) in roots tissue (around 17.23 ± 3.12 RPKM), suggesting that this gene could be responsible for the dense root system and tillering potential described in *S. spontaneum*.

4. Conclusions

Even with large advances in the sequencing technologies, the assembly of complex genomes still represents a bottleneck, mainly due to polyploidy and high heterozygosity. The development of new bioinformatics efforts can contribute to overcoming these constraints, especially, for complete genomes of the closely related organism, in which the methods based on reference assembly can be applied. Using the PGA pipeline, we provided a high-quality assembly of gene regions in the *T. aestivum* and *S. spontaneum*, proving that PGA can be more efficient than conventional genome assemblers in cases of

Table 4. Summary of the *S. spontaneum* promoter regions (≥2,000 bp upstream from the gene) classified by gene expression profile and intensity

		Gene expression profile						
		Constitutive	Tissue specific					
			Leaf	Root	Node	Internode	Bud	Apical meristem
Gene expression intensity	Weak	80	33	20	4	0	2	1
	Strong	7	25	2	1	0	0	2

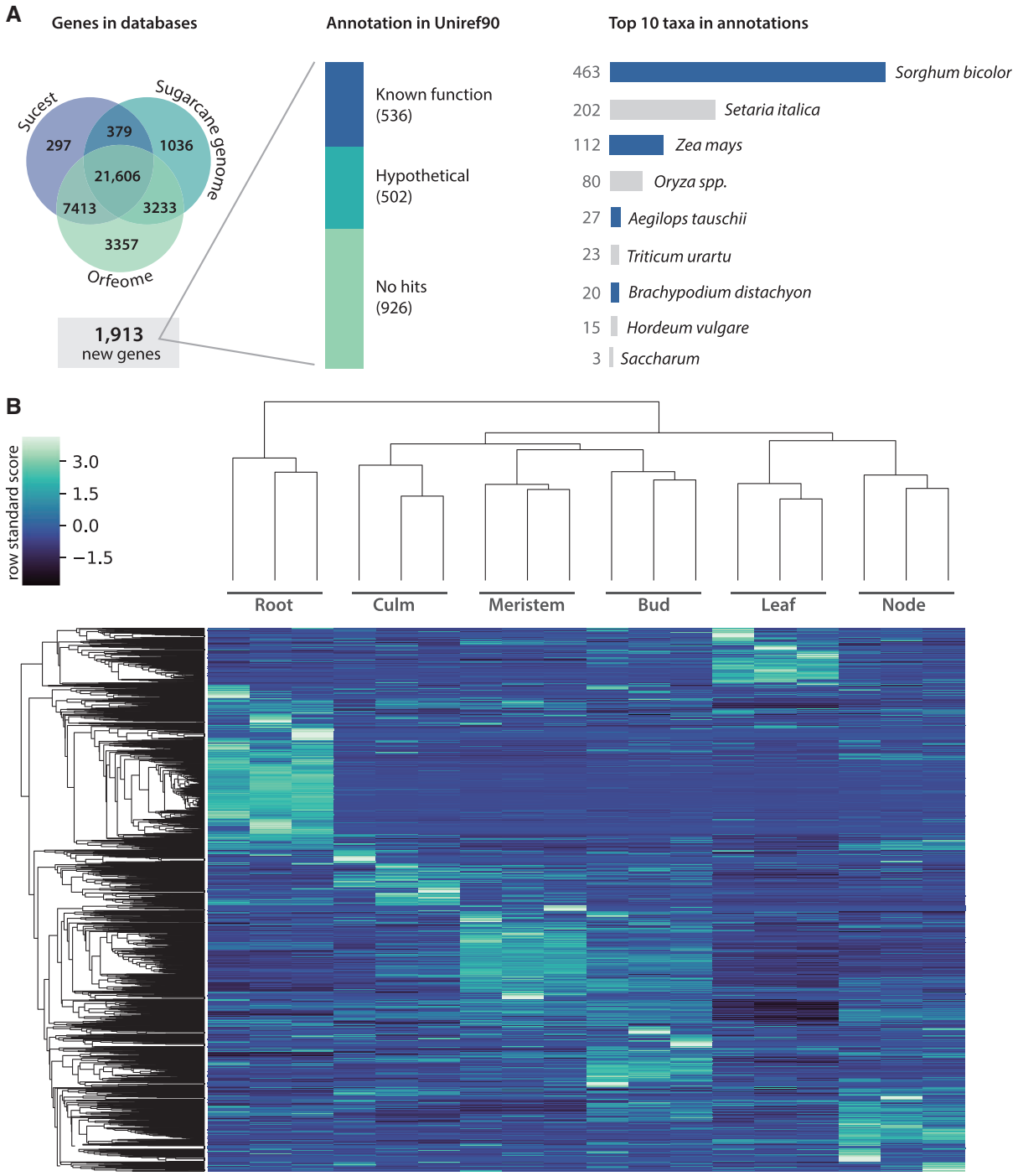


Figure 4. New contributions for sugarcane databases. (A) Results from the comparison of *Saccharum spontaneum 1.0* assemble against Sugarcane databases (Gene prediction from sugarcane genome, SUCEST-FUN and ORFeome) and (B) the expression pattern of new identified genes in different tissues.

complex genomes and using low coverage DNA sequencing. The low memory requirement by PGA in comparison with *de novo* assembly strategy is also an advantage. Our findings of *S. spontaneum* genome highlighted for the first time the molecular basis of some noteworthy features of this biomass, like the high productivity and the resistance face biotic and abiotic stress. Those results can be employed in future functional and genetic studies beyond supporting the development of new varieties of sugarcane for the agronomic industry.

Availability of data

The DNA and RNA Illumina reads were submitted to SRA (Sequence Read Archive) from NCBI under BioProject accession number PRJNA474618. Also, *S. spontaneum* genome and gene prediction are available to download at the Saccharum database (http://www.lge.ibi.unicamp.br/spontaneum - date last accessed 25 January 2019) constructed using the scripts provided by EUCANEXT.[69]

Table 5. Probable symbiosis-related genes expressed in *S. spontaneum* roots

Gene ID	Protein/function	Expression (RPKM)	Reference
SS43844	MATH domain protein	6.33	60
SS34074	LRR receptor	6.66	60
SS37878		2.88	
SS50399		2.04	
SS37936	LysM domain	2.81	60
SS50999	Calcium and calmodulin-dependent protein kinase	4.59	61
SS11486	Salt stress-induced proteins	19.09	62
SS38841	Jacalin-like lectin	20.13	63
SS32869		16.06	
SS35410		2.38	
SS41259	Glutathione S-transferase	1.31	62
SS15872	WRKY transcription factor	3.56	64,65
SS42099		2.01	
SS35024	Blue copper binding proteins	23.70	62
SS38176	Lipid transfer protein	210.90	61
SS33290	Arabinogalactan	10.50	62
SS48435	Pectinesterase	2.92	62
SS46591		1.09	
SS47675	Xyloglucan fucosyltransferase	2.13	62

Acknowledgements

We are also thankful to (i) Life Sciences Core Facility (LaCTAD), University of Campinas (UNICAMP) for the computational infrastructure and (ii) GranBio Investimentos S.A. for *S. spontaneum* accessions.

Funding

We thank the Brazilian National Council for Scientific and Technological Development (CNPq–RHAE, grants number 350474/2013-3 and 35081/2015-1), the Sao Paulo Research Foundation (FAPESP, grant numbers 2014/09638-0 and 2012/05890-1) and Center for Computational Engineering and Sciences—FAPESP/Cepid (2013/08293-7) for the financial support and scholarships.

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Gregory, T.R., Nicol, J.A. and Tamm, H. 2007, Eukaryotic genome size databases, *Nucleic Acids Res*., **35**, D332–8.
2. Gore, M.A., Chia, J.-M., Elshire, R.J., et al. 2009, A First-Generation Haplotype Map of Maize, *Science*, **326**:1115–1117.
3. Schatz, M.C., Witkowski, J. and McCombie, W.R. 2012, Current challenges in *de novo* plant genome sequencing and assembly, *Genome Biol*., **13**, 243.
4. International Wheat Genome Sequencing Consortium (IWGSC). 2014, A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome, *Science*, **18**, 1251788.
5. Nystedt, B., Street, N.R., Wetterbom, A., et al. 2013, The Norway spruce genome sequence and conifer genome evolution, *Nature*, **497**, 579–84.
6. Kim, J., Larkin, D.M., Cai, Q., et al. 2013, Reference-assisted chromosome assembly, *Proc. Natl. Acad. Sci. U.S.A*., **110**, 1785–90.
7. D'Hont, A. and Glaszmann, J.C. 2001, Sugarcane genome analysis with molecular markers: a first decade of research, *Proc. Int. Soc. Sugar Cane Technol*., **24**, 556–9.
8. Garsmeur, O., Charron, C., Bocs, S., et al. 2011, High homologous gene conservation despite extreme autopolyploid redundancy in sugarcane, *New Phytol*., **189**, 629–42.
9. Grivet, L. and Arruda, P. 2002, Sugarcane Genomics: Depicting the Complex Genome of an Important Tropical Crop, *Curr. Opin. Plant Biol*., **5**, 122–127.
10. Carvalho-Netto, O.V., Bressiani, J.A., Soriano, H.L., et al. 2014, The Potential of the Energy Cane as the Main Biomass Crop for the Cellulosic Industry, *Chem. Biol. Technol. Agric*., **1**:20.
11. Sreenivasan, T.V. and Ahloowalia, B.S. 1987, Cytogenetics. In: Heinz D.J., ed. *Sugarcane Improvement through Breeding*, pp. 211–53. Elsevier: Amsterdam.
12. Nishiyama, M.Y. Jr, Vicente, F.F.R., Lembke, C.G., et al. 2010, The SUCEST-FUN regulatory network database: designing an energy grass, *Proc. Int. Soc. Sugar Cane Technol*., **27**, 1–10.
13. Nishiyama, M.Y. Jr, Ferreira, S.S., Tang, P.-Z., Becker, S., Pörtner-Taliana, A. and Souza, G.M. 2014, Full-length enriched cDNA libraries and ORFeome analysis of sugarcane hybrid and ancestor genotypes, *PLoS One*, **9**, e107351.
14. Miller, J.R., Dilley, K.A., Harkins, D.M., et al. 2017, Initial genome sequencing of the sugarcane CP 96-1252 complex hybrid, *F1000Res*., **6**, 688.
15. Garsmeur, O., Droc, G., Antonise, R., et al. 2018, A mosaic monoploid reference sequence for the highly complex genome of sugarcane, *Nat. Commun*., **9**, 2638.
16. Zhang, J., Zhang, X., Tang, H., et al. 2018, Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L., *Nat. Genet*., **50**: 1565–1573.
17. Zeng, Y. and Yang, T. 2002, RNA isolation from highly viscous samples rich in polyphenols and polysaccharides, *Plant Mol. Biol. Rep*., **20**, 417.
18. Le Provost, G., Herrera, R., Paiva, J.A., Chaumeil, P., Salin, F. and Plomion, C. 2007, A micromethod for high throughput RNA extraction in forest trees, *Biol. Res*., **40**, 291–7.
19. Carvalho, C.R., Clarindo, W.R., Praça, M.M., Araújo, F.S. and Carels, N. 2008, Genome size, base composition and karyotype of *Jatropha curcas* L., an important biofuel plant, *Plant Sci*., **174**, 613–7.
20. Langmead, B. and Salzberg, S.L. 2012, Fast gapped-read alignment with Bowtie 2, *Nat. Methods*, **9**, 357–9.
21. Grabherr, M.G., Haas, B.J., Yassour, M., et al. 2011, Full-length transcriptome assembly from RNA-Seq data without a reference genome, *Nat. Biotechnol*., **29**, 644–52.
22. Xue, W., Li, J.-T., Zhu, Y.-P., et al. 2013, L_RNA_scaffolder: scaffolding genomes with transcripts, *BMC Genomics*, **14**, 604.
23. Kent, W.J. 2002, BLAT—the BLAST-like alignment tool, *Genome Res*., **12**, 656–64.
24. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. and Pirovano, W. 2011, Scaffolding pre-assembled contigs using SSPACE, *Bioinformatics*, **27**, 578–9.
25. Beier, M.S., Himmelbach, A., Colmsee, C., et al. 2017, Construction of a map-based reference genome sequence for barley, *Hordeum vulgare L*., *Sci. Data*, **4**, 170044.
26. Goodstein, D.M., Shu, S., Howson, R., et al. 2012, Phytozome: a comparative platform for green plant genomics, *Nucleic Acids Res*., **40**, D1178–86.
27. Slater, G.S.C.M. and Birney, E. 2005, Automated generation of heuristics for biological sequence comparison, *BMC Bioinform*., **6**, 31.
28. Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. 2006, AUGUSTUS: *ab initio* prediction of alternative transcripts, *Nucleic Acids Res*., **34**, W435–9.

29. Quinlan, A.R. and Hall, I.M. 2010, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**, 841–2.
30. Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R. and Wu, C.H. 2007, UniRef: comprehensive and non-redundant UniProt reference clusters, *Bioinformatics*, **23**, 1282–8.
31. Paterson, A.H., Bowers, J.E., Bruggmann, R., et al. 2009, The *Sorghum bicolor* genome and the diversification of grasses, *Nature*, **457**, 551–6.
32. Schnable, P.S., Ware, D., Fulton, R.S., et al. 2009, The B73 maize genome: complexity, diversity, and dynamics, *Science*, **326**, 1112–5.
33. Wang, J., Zhang, G., Liu, X., et al. 2012, Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential, *Nat. Biotechnol*., **30**, 549–54.
34. Li, R., Zhu, H., Ruan, J., et al. 2010, *De novo* assembly of human genomes with massively parallel short read sequencing, *Genome Res*., **20**, 265–72.
35. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. 2013, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, *Genome Biol*., **14**, R36.
36. Bairoch, A. and Apweiler, R. 2000, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res*., **28**, 45–8.
37. Ouyang, S., Zhu, W., Hamilton, J., et al. 2007, The TIGR rice genome annotation resource: improvements and new features, *Nucleic Acids Res*., **35**, D883–7.
38. Lamesch, P., Berardini, T.Z., Li, D., et al. 2012, The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools, *Nucleic Acids Res*., **40**, D1202–210.
39. Cardoso-Silva, C.B., Costa, E.A., Mancini, M.C., et al. 2014, *De novo* assembly and transcriptome analysis of contrasting sugarcane varieties, *PLoS One*, **9**, e88462.
40. Haas, B.J., Delcher, A.L., Mount, S.M., et al. 2003, Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies, *Nucleic Acids Res*., **31**, 5654–66.
41. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, **31**, 3210–12.
42. Conesa, A. and Götz, S. 2008, Blast2GO: a comprehensive suite for functional analysis in plant genomics, *Int. J. Plant Genomics*, **2008**, 619832.
43. Initiative, I.B. 2010, Genome sequencing and analysis of the model grass *Brachypodium distachyon*, *Nature*, **463**.
44. Peng, Z., Lu, Y., Li, L., et al. 2013, The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*), *Nat. Genet*., **45**, 456–61.
45. Li, L., Stoeckert, C.J. and Roos, D.S. 2003, OrthoMCL: identification of ortholog groups for eukaryotic genomes, *Genome Res*., **13**, 2178–89.
46. Katoh, K. and Standley, D.M. 2013, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Mol. Biol. Evol*., **30**, 772–80.
47. Stamatakis, A. 2014, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, *Bioinformatics*, **30**, 1312–13.
48. Librado, P., Vieira, F.G. and Rozas, J. 2012, BadiRate: estimating family turnover rates by likelihood-based methods, *Bioinformatics*, **28**, 279–81.
49. Trapnell, C., Williams, B.A., Pertea, G., et al. 2010, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, *Nat. Biotechnol*., **28**, 511–15.
50. Li, B. and Dewey, C.N. 2011, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, *BMC Bioinformatics*, **12**, 323.
51. Kong, L., Zhang, Y., Ye, Z.-Q., et al. 2007, CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine, *Nucleic Acids Res*., **35**, W345–9.
52. Yanai, I., Benjamin, H., Shmoish, M., et al. 2005, Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification, *Bioinform. Oxf. Engl*., **21**, 650–9.
53. Zerbino, D.R. and Birney, E. 2008, Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs, *Genome Res*., **18**, 821–9.
54. Arendsee, Z.W., Li, L. and Wurtele, E.S. 2014, The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*, *GigaScience*, **6**, 1–7.
55. Dufayard, J.-F., Bettembourg, M., Fischer, I., et al. 2017, New insights on leucine-rich repeats receptor-like kinase orthologous relationships in angiosperms, *Front. Plant Sci*., **8**:381.
56. Hirsch, S. and Oldroyd, G.E.D. 2009, GRAS-domain transcription factors that regulate plant development, *Plant Signal. Behav*., **4**, 698–700.
57. Garcia Tavares, R., Lakshmanan, P., Peiter, E., et al. 2018, ScGAI is a key regulator of culm development in sugarcane, *J. Exp. Bot*., **69**(16): 3823–3837.
58. Schneider, R., Hanak, T., Persson, S. and Voigt, C.A. 2016, Cellulose and callose synthesis and organization in focus, what's new? *Curr. Opin. Plant Biol*., **34**, 9–16.
59. Falter, C., Zwikowics, C., Eggert, D., et al. 2015, Glucanocellulosic ethanol: the undiscovered biofuel potential in energy crops and marine biomass, *Sci. Rep*., **5**, 13722.
60. Shrivastava, N., Jiang, L., Li, P., et al. 2018, Proteomic approach to understand the molecular physiology of symbiotic interaction between *Piriformospora indica* and *Brassica napus*, *Sci. Rep*., **8**, 5773.
61. Miller, J.B., Pratap, A., Miyahara, A., et al. 2013, Calcium/calmodulin-dependent protein kinase is negatively and positively regulated by calcium, providing a mechanism for decoding calcium responses during symbiosis signaling, *Plant Cell*, **25**, 5053–66.
62. Hohnjec, N., Vieweg, M.F., Pühler, A., Becker, A. and Küster, H. 2005, Overlaps in the transcriptional profiles of *Medicago truncaluta* roots inoculated with two different Glomus fungi provide insights into the genetic program activated during arbuscular mycorrhiza, *Plant Physiol*., **137**, 1283–301.
63. Fiorilli, V., Vallino, M., Biselli, C., Faccio, A., Bagnaresi, P. and Bonfante, P. 2015, Host and non-host roots in rice: cellular and molecular approaches reveal differential responses to arbuscular mycorrhizal fungi, *Front. Plant Sci*., **6**, 636.
64. Zhang, C., Wang, D., Yang, C., et al. 2017, Genome-wide identification of the potato WRKY transcription factor family, *PLoS One*, **12**, e0181573.
65. Pii, Y., Astegno, A., Peroni, E., Zaccardelli, M., Pandolfini, T. and Crimi, M. 2009, The *Medicago truncaluta* N5 gene encoding a root-specific lipid transfer protein is required for the symbiotic interaction with *Sinorhizobium meliloti*, *Mol Plant Microbe Interact*, **22**(12), 1577–87.
66. Mirshad, P.P. and Puthur, J.T. 2017, Drought tolerance of bioenergy grass *Saccharum spontaneum* L. enhanced by arbuscular mycorrhizae, *Rhizosphere*, **3**, 1–8.
67. Battaglia, M. and Covarrubias, A.A. 2013, Late Embryogenesis Abundant (LEA) proteins in legumes, *Front. Plant Sci*., **4**, 190.
68. Li, Y., Chen, X., Chen, Z., Cai, R., Zhang, H. and Xiang, Y. 2016, Identification and expression analysis of BURP domain-containing genes in *Medicago truncatula*, *Front. Plant Sci*., **7**, 485.
69. Nascimento, L.C., Salazar, M.M., Lepikson-Neto, J., et al. 2017, EUCANEXT: an integrated database for the exploration of genomic and transcriptomic data from *Eucalyptus* species, *Database J. Biol. Databases Curation*, **2017**:doi.org/10.1093/database/bax079.