Centro Nacional de Pesquisa em Energia e Materiais
Laboratório Nacional de Ciência e Tecnologia do Bioetanol

MeT 02/2016

Surveying the complex polyploid sugarcane genome sequence using synthetic long reads

Diego Mauricio Riaño-Pachón
Lucia Mattiello
Larissa Prado da Cruz

Abstract

Today there is a lack of publicly available genomics data for sugarcane; availability of such information will eventually strengthen and accelerate breeding programs. This lack of information is partly due to sugarcane's complex genome structure, which is not amenable to current high-throughput short-read sequencing technologies. Current sugarcane cultivars are interspecific hybrids, with a ploidy level between 8 and 12, and with an estimated haploid genome size of approx. 760-930Mbps. We have used the new TruSeq Synthetic Long Read sequencing technology from Illumina, in a pilot project, to sequence the sugarcane genome (variety SP80-3280) at a shallow coverage. We have generated 9 libraries, accounting for more than 5Gbp of sequence data, thus giving an estimated coverage of around 5x of the haploid genome. The current assembly has over 1Gbp of assemble long reads, and we could annotate 300,000 protein-coding genes exploiting RNASeq data previously generated in the group. Identification of a highly conserved gene set in eukaryotes, have revealed a coverage of approx. 90% of the gene space. The genome assembly, gene prediction and further data are available via http://bce.bioetanol.cnpem.br/sugarcanegenome. We have released a BLAST server to access the draft genome sequence, available at: http://bce.bioetanol.cnpem.br/ctbeblast.

Keywords: synthetic long reads, sugarcane, genome, annotation, transcription factors, orthologues, natural antisense transcripts, tRNAs

Table of contents

Abstract 2
Introduction 5
Materials and methods 7
Plant material and DNA extraction 7
TruSeq synthetic long read library preparation and sequencing 8
Characteristics of TruSeq synthetic long reads 8
Genome assembly and gene prediction 9
Results 10
Characteristics of TruSeq synthetic long reads 10
Species representativeness among truseq synthetic long reads. 11
De novo genome assembly. 12
Completeness of the de novo genome assembly 14
Identification of repeat sequences and Gene prediction 15
Genome data availability 19
Author contributions 19
References 20
Appendix 23
Appendix A. "Reserved analyses" notice for access to the genome data generated at CTBE 23

List of Figures

Figure 1. Length distribution of synthetic long reads 10
Figure 2. Profile mismatch resulting from comparing synthetic long reads with sugarcane BAC sequences available at the NCBI..... 11
Figure 3. Taxonomic assignment of TruSeq Synthetic Long reads. 12
Figure 4. Identification and classfication of TAPs in siz Angiosperm species (ATHA: Arabidopsis thaliana, OSAT: Oriza sativa, SITA: Setaria itálica, SBIC: Sorghum bicolor, ZMAY: Zea mays: SACC: Saccharum SP80-3280) 17
Figure 5. Groups of orthologues shared among the angiosperms Sugarcane SP80-3280 (SACC) Arabidopsis thaliana (ATHA), Oryza sativa (OSAT), Setaria italica (SITA), Zea mays (ZMAY), and Sorghum bicolor (SBIC). Venn diagram generated with http://www.interactivenn.net/ 18
Figure 6. Number of tRNA genes for the different aminoacids and anti-codons 19

List of Tables

Table 1. Summary of sequencing throughput per library..... 10
Table 2. Settings used for the de novo genome assembly using WGS v8.2 (*= default setting) 13
Table 3. Summary statistics of the de novo genome assemblies. 14
Table 4. Frequency of repeat sequences in the draft assembly..... 15
Table 5. Rounds of genome annotation and statistics. 16
Table 6. Weights for the difference types of evidence used to generate a high quality set of protein-coding genes using EVM 17
Table 7. Statistics about the clustering of orthologuous genes and exclusive proteins 18

Introduction

Sugarcane is an economically important crop used as source of sugar and ethanol (MAPA 2012), electricity generation (HASSUANI *et al.* 2005) and, more recently, used to produce other bio-products, e.g., paper (CHANDEL *et al.* 2012). The world is moving towards an increased use of renewable sources of energy, while increasing its energy demands. However, to supply this energy demands, the production of sugarcane should increase without compromising other land uses (VALDES 2011), thus requiring new and better varieties. The development of high throughput molecular technologies ('omics', and especially genomics) in the last decades is opening new ways to approach the study of important crops and their improvement in a time and cost-effective way (JACKSON *et al.* 2011). Leveraging the information in the genome can speed up breeding programs and genetic engineering. However, despite its economic importance, the sugarcane genome has not been deciphered yet.

Sugarcane has a haploid genome of relatively small size (~1Gpb), however, modern sugarcane cultivars are polyploids derived from interspecific hybridization between *S. officinarum* L. and *S. spontaneum* L. (ROACH 1969) reaching up to 130 chromosomes distributed among ~12 homologous groups (D'HONT 2005; GRIVET and ARRUDA 2002) with a total genome size reaching 10Gpb (LE CUNFF *et al.* 2008). Polyploidy is an important mechanism for the evolution and diversification for both plants and animals (DUFRESNE *et al.* 2014), but it hampers genome sequencing, assembly and annotation due to the usually high levels of polymorphism across the homeologous chromosomes, and the large proportion of repetitive sequences (SOUZA *et al.* 2011; YANG *et al.* 2011). Unfortunately, sugarcane does not have diploid progenitors to allow faster and easier genome assembly (GARCIA *et al.* 2013) in the same fashion as made for potato (XU *et al.* 2011), cotton (LI *et al.* 2014) and banana (D'HONT *et al.* 2012).

Nevertheless, several sugarcane genome projects are undergoing around the world, e.g., South Africa, Australia, France. In Brazil the main effort is under the umbrella of the BIOEN-FAPESP project that started in 2008 (http://www.fapesp.br/bioen; http://agencia.fapesp.br/pontape_inicial/10838/). There are two main approaches being followed: one using a collection of BACs enriched with genic regions, and a whole genome shotgun approach. Recently a collection of

317 BAC sequences were released to the public (DE SETTA *et al.* 2014). Furthermore, an additional attempt to sequence gene-enriched regions exploiting methyl-filtration also made available genome data, although with a high level of fragmentation that limits its applicability in functional genomics studies (GRATIVOL *et al.* 2014). With that scenario, there is no envisioned date for the release of a complete or close to complete sugarcane genome sequence, and release of much appreciate preliminary data from the BIOEN project has been scarce. Thus, interested researchers have to turn to other resources. One of the largest publicly available sequence resources is a collection of almost 300.000 Expressed Sequence Tags (EST) deposited in the GenBank. This is the largest collection of sugarcane sequences available, and most of them were generated by researchers in Brazil (MENOSSI *et al.* 2008). More recently, traditional and new sequencing technologies have been employed to identify microRNAs (miRNAs) in sugarcane, and the resulting datasets have been made available to the public (CARNAVALE BOTTINO *et al.* 2013; FERREIRA *et al.* 2012; GENTILE *et al.* 2013; ORTIZ-MOREA *et al.* 2013; VICENTINI *et al.* 2015; ZANCA *et al.* 2010), but it appears that the identification of sugarcane miRNAs is far from saturation, when compared with surveys in sorghum or rice cf. miRBase (http://www.mirbase.org/; (KOZOMARA and GRIFFITHS-JONES 2011)). Nowadays, it is clear that the development of next generation sequencing technologies using short reads, e.g., Illumina, together with the development of new, efficient and precise bioinformatics algorithms, is paving the way towards a comprehensive knowledge of the transcriptome landscape of species of commercial and social interest whose genome is not yet available (e.g., (GUPTA *et al.* 2013; LULIN *et al.* 2012; YAZAWA *et al.* 2013)), Recently, such approaches have been applied to the study of the sugarcane transcriptome in different conditions or of different genotypes (CARDOSO-SILVA *et al.* 2014; NISHIYAMA *et al.* 2014), as well as a reduced representation of the genome sequence, enriched in genic regions, obtained by methyl filtration (GRATIVOL *et al.* 2014). Although these efforts have been made, lack of publicly available genomics information in a structured and amicable way, is delaying the progress in several fronts in the sugarcane scientific and breeding landscapes. For instance, no genetically modified sucarcane has been released despite some reported field trials (CHEAVEGATTI-GIANOTTO *et al.* 2011; HOTTA *et al.* 2010).

Even more, novel uses of current bioinformatics algorithms are able to resolve allele specific sequences; this is of the utmost importance for crops with high ploidy and polymorphic levels, such as sugarcane [18]. We have employed RNASeq on different stages of sugarcane leaf development to develop a strand-specific transcriptome that is publicly available (MATTIELLO *et al.* 2015). However, such dataset is still very limited. For example, there are not promoter sequences, and in many cases, the assembled contigs/transcripts only represent a part of the whole transcript, just to name a couple of issues. These limits could be greatly overcome with the availability of genome data.

Here we used the new Illumina TruSeq Synthetic Long Read sequencing technology (ILLUMINA 2015; MCCOY *et al.* 2014) to survey the sugarcane genome (cultivar SP80-3280) using 9 libraries, each generating approx. 600Mbps, thus giving an estimated coverage between 4 and 5 of the monoploid genome. The generated long reads and their assembly will provide useful information for functional genomics studies.

Materials and methods

Plant material and DNA extraction

Sugarcane stalks from genotype SP80-3280 (*Saccharum* spp.) were gently provided by Centro de Tecnologia Canavieira (CTC), Piracicaba SP, Brazil. Stalks were sectioned in order to contain only one bud and were germinated in trays containing vermiculite. After one month, plants with homogenous phenotype were transferred to pots (3.5 L) containing pine-bark substrate and vermiculite (1:1). Plants were fertilized with N:P:K every 15 days. The pots were watered daily. The leaf roll of a two-month old plant was collected and immediately frozen on liquid nitrogen. Tissue was grounded until fine powder using a mortar and pestle. High molecular weight DNA was extracted using CTAB as follows: 100 mg of fresh frozen tissue was added to 600 µL of heated (65°) extraction buffer corrected to pH 5 (100 mM Tris-HCl 1M pH 8, 20 mM EDTA pH 8; 1,4 M NaCl; 2% CTAB; 1% PVP; 0,1 % β-mercaptoethanol), incubated for 60 minutes with eventual shaking. One volume of chloroform:isoamyl alcohol (24:1) was added, homogenized and centrifuged for 10 minutes at 14.000 rpm. The supernatant was transferred to a new tube and 0.7 V of isopropanol and 1/10 V of sodium acetate (3 M) was added. After overnight incubation (-20°C), the mixture was centrifuged for 10 minutes at

14.000 rpm. The supernatant was discarded and the pellet was washed with ice-cold 70% alcohol. The DNA was resuspended in 20 µL of TE buffer, treated with RNAse, visualized in 1% agarose gel and quantified in a spectrophotometer.

TruSeq synthetic long read library preparation and sequencing

Six µg of DNA were sent to Illumina Inc, CA for DNA sequencing using TruSeq Synthetic long reads technology, through their FastTrack Sequencing Service. Briefly, genomic DNA is sheared into long fragments, ligated to amplification adapters and separated into 384-well plates using limiting dilution, such that each well would contain approximately 300 molecules (McCoy *et al.* 2014). The fragments in each well are amplified by long-range PCR, and later used to make Nextera libraries (Caruccio 2011). The fragments of different wells will have different barcode indexes, introduced during the library preparation. Fragments are sequenced in an Illumina HiSeq2000 system using paired-end chemistry. Sequenced paired-end reads are demultiplexed and assembled separately by barcode using a proprietary short read assembler (McCoy *et al.* 2014). Sample preparation kits are available from Illumina and the assembly software is accessible on BaseSpace using the TruSeq Long-Read Assembly App.

Characteristics of TruSeq synthetic long reads

In order to evaluate the quality of the TruSeq synthetic long reads, they were aligned to the collection of 352 BAC sugarcane sequences available at NCBI, most of them originating from genotype R570. Reads were aligned to the reference set using BLAT (Kent 2002). BLAT results were filtered with an *in-house* Perl script that extracted the alignment positions and the corresponding sequences, and then performed a Needleman and Wunsch global alignment (Needleman and Wunsch 1970) with the tool `stretcher` from the EMBOSS suite (Myers and Miller 1988; Olson 2002). The proportion of mismatches along the alignment were registered and expressed as a per cent of the alignment length. Alignments with more than 10% of the read length represented as gaps were omitted from the mismatch analysis.

In order to evaluate possible contamination of the synthetic long reads, they were compared against the NCBI's nucleotide database (nt) using BLAST (Altschul *et al.* 1990), keeping only the best hit for each read with an e-value smaller than

1e-5. Results were imported into a MySQL database and visualized in R (R CORE TEAM 2015).

Genome assembly and gene prediction

Only synthetic long reads longer than 1.5Kbp and with BLAST hits against sequences from Viridiplantae were used for genome assembly. Prior to assembly, reads originating from the mitochondria and the chloroplast were excluded using `mirabait` (CHEVREUX *et al.* 1999), taking as references ACC: NC_008360.1 and ACC: NC_005878.2, respectively. Reads putatively originating from the nuclear plant genome were assembled using Celera's Whole-Genome Shotgun Assembler v 8.2, following the recommendations in (MCCOY *et al.* 2014) for the assembly of synthetic long reads. The assembled sequences corresponding to scaffolds, singletons and degenerates, were joined, and a non-redundant assembly was created using CD-HIT (FU *et al.* 2012), merging 100% identical sequences and sub-sequences, the resulting assembly was used in further steps.

CEGMA v2.5 and BUSCO were employed to detect a set of highly conserved genes in eukaryotic genomes and to evaluate the completeness of the assembly (PARRA *et al.* 2007; SIMAO *et al.* 2015).

Prior to gene prediction, we identified and masked sequence repeats using RepeatMasker v4.0.5 (http://www.repeatmasker.org). Then, the gene models identified by CEGMA, see above, were used to train the *ab initio* gene predictor Augustus (STANKE *et al.* 2006). 380 million strand-specific RNASeq short paired-end reads obtained from plants of the same cultivar (MATTIELLO *et al.* 2015) were exploited to provide experimental evidence for gene prediction using BRAKER1 (HOFF *et al.* 2015) and Augustus (STANKE *et al.* 2006).

We employed ITSx to identify the ribosomal RNA genes and tRNA-scan-SE v1.3 to identify tRNAs. We plan to exploit public miRNA and sRNA data (ORTIZ-MOREA *et al.* 2013) in the near future to improve the genome annotation.

Among the protein-coding genes we are particularly interested in these involved in the regulation of transcription, as they could be targets for modification with the goal of obtaining desired phenotypes. We have applied the methods described in (PEREZ-RODRIGUEZ *et al.* 2010) in order to identify genes belonging to transcription factor families and families of other transcriptional

regulators, this in the draft genome sequence of sugar cane as well as other plant species.

Results

Characteristics of TruSeq synthetic long reads

A total of nine TruSeq synthetic long read libraries, sequenced in two batches, were produced by Illumina Inc., yielding 1,378,917 reads longer than 1.5Kbp, or 5,642,855,018 bases (Table 1). The underlying 1,966,604,928 short reads amount to 393,320,985,600bp, which would translate to an estimated coverage of 393x of the haploid genome. The maximum read length achieved was 20,918 bp, the length profile of the synthetic long reads is show in Figure 1**Error! Reference source not found.**. 36% percent of the reads were longer than 4.5 Kbp.

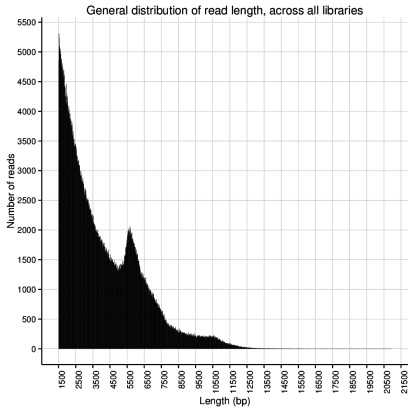


Figure 1. Length distribution of synthetic long reads

Table 1. Summary of sequencing throughput per library

Batch	Library ID	Total sequence (Mbp) reads > 1.5Kbp	Number of reads > 1.5Kbp	Number of reads > 4.5Kbp	Number of reads > 6Kbp	Max read length (bp)
1	LRAAD-01	484,733,291	111,523	56,622	22,345	14,201
1	LRAAD-04	511,210,351	116,592	60,151	24,063	12,528
1	LRAAD-05	488,831,968	110,652	58,233	23,213	12,932
2	LRAAD-06	584,402,404	147,59	45,502	26,978	20,918
2	LRAAD-07	601,548,339	153,013	46,379	27,118	19,368
2	LRAAD-08	707,299,521	172,361	57,149	34,733	20,799
2	LRAAD-12	671,285,370	174,982	51,052	28,464	18,944
2	LRAAD-13	915,082,750	220,464	75,055	46,013	18,98
2	LRAAD-14	678,461,024	171,74	53,125	31,051	18,626
TOTAL		**5,642,855,018**	**1.378.917**	**503.268**	**263.978**	

The Figure 2 shows the rate of mismatches between BAC sequences and the synthetic long reads. There is an approximately uniform rate of mismatches along the read, of approx. 8%). It is important to note here that most of the BAC sequences come from the French cultivar R570, and thus the mismatch rate is to a large extend a measure of the polymorphism between two Brazilian and the French cultivars. The red and blue lines in Figure 2 shows the expected number of mismatches arising only from polymorphisms (i.e., without sequencing or assembly errors) for coding regions, using two estimates, i.e., one SNP every 50pb of coding sequence (CORDEIRO *et al.* 2006) and one every 86bp (CARDOSO-SILVA *et al.* 2014), both are underestimates of the genome-wide polymorphism rate, the actual rate would be between 2 and 8% per position.

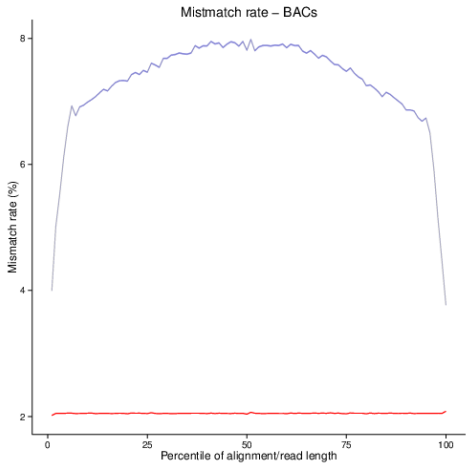


Figure 2. Profile mismatch resulting from comparing synthetic long reads with sugarcane BAC sequences available at the NCBI.

Species representativeness among truseq synthetic long reads.

Figure 3 shows the results of comparing the long reads to the NCBI nucleotide database. 88.7% of the reads can be assigned to the Eukaryota superkingdom, the few remaining years either cannot be mapped to the database or map to other superkingdoms (Figure 3A). Most of these 1.224.061 reads, originate from the green plant plan lineage (99.97%, Figure 3B), where the all the genus represented with more than 1% of the reads belong to the grasses, and the read with hits to *Sorghum* and *Saccharum* represent more than 80%. (Figure 3C). The reads with hits to Viridiplantae were kept for the *de novo* genome assembly.

De novo genome assembly.

Only reads originating from Viridiplantae (see above) were kept for genome assembly. Currently there are two main software families for de novo genome assembly. Software based on the de Bruijn graph and software based on the overlap-layout-consensus approach. The former was developed to specifically deal with large amount of short reads, and the later was the approach originally developed to assemble the human genome sequence, when the number of sequencing reads is not very large, reads are long, and of very high quality (Li *et al.* 2012). The data generated with the TruSeq Synthetic Long Reads approach can be assembled using the OLC approach (McCoy *et al.* 2014). We tried two of the main software packages implementing variations of the OLC approach, i.e, MIRA v4.9 (Chevreux *et al.* 1999) and WGS v8.2 (Myers *et al.* 2000), we tried both packages with WGS resulting in a better behaviour; in the following we will only show the results obtained with that assembler.

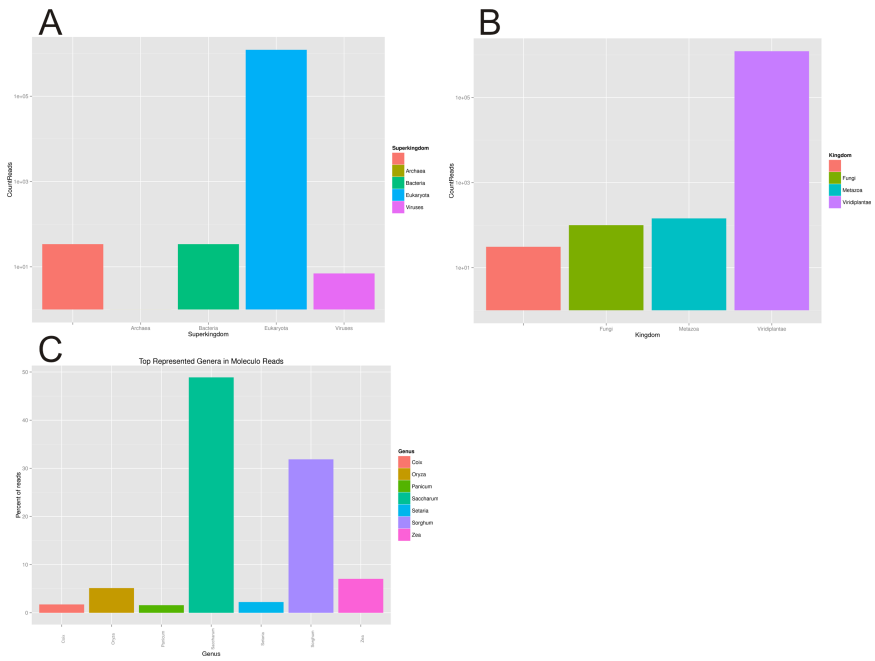


Figure 3. Taxonomic assignment of TruSeq Synthetic Long reads.

For the assembly with WGS v8.2 we used similar parameters as those described in (McCoy *et al.* 2014), except for some of the error parameters that were left in their default settings, as they would be more useful for polymorphic genomes, such as sugarcane (

Table 2).

Table 2. Settings used for the de novo *genome assembly using WGS v8.2 (*= default setting)*

Setting	Value
unitigger	bogart
merSize	31
ovlMinLen	100
ovlErrorRate	0.06*
cnsErrorRate	0.06*
cgwErrorRate	0.1*
utgGraphErrorRate	0.03*
utgGraphErrorLimit	3.25*
utgMergeErrorRate	0.045*
utgMergeErrorLimit	5.25*

WGS produces contigs, singletons and degenerate sequences. Singletons are reads that could not be assembled, but as we have long reads these singletons carry important genome information and will be kept. Contigs, are overlapping collection of unitigs, which in turn are assemblies of overlapping reads, this is the main output of an assembler. Last, the degenerates, are unitigs that could not be used to form contigs, they are assemblies of overlapping reads, we are also keeping these for further analyses. We took all contigs, singletons and degenerate sequences and this constitutes our preliminary assembly. This preliminary assembly resulted in 202.892 sequences (contigs + singletons + degenerates, in the following called just contigs), with a total length of over 1.1 Gbp. Examining the preliminary assembly we recognized that there was still some level of redundancy. CD-HIT was used to eliminate the redundancy in the preliminary assembly, merging sequences that were 100% identical. Table 3 shows summary statistics of both assemblies, and also for the Methyl-filtration assembly from (GRATIVOL *et al.* 2014) for comparison. The data in the table clearly show that using synthetic long reads clearly increases several of the metrics, in general increasing the contiguity of the genome reconstruction.

Table 3. Summary statistics of the de novo *genome assemblies.*

Assembly	Original Assembly	Non-redundant assembly	Methyl-filtration (GRATIVOL *et al.* 2014)
contigs (>= 0 bp)	202,892	199,028	1,109,444
contigs (>= 1000 bp)	182,296	181,569	146,177
Total length (>= 0 bp)	1,172,338,336	1,169,948,913	674,032,540
Total length (>= 1000 bp)	1,162,215,769	1,160,325,082	346,121,853
Largest contig	115,913	115,913	35,917
GC (%)	43.04	43.04	48
N50	8,438	8,451	2,083
N75	5,188	5,203	1,636
N's per 100 kbp	0	0	46,984

Completeness of the de novo genome assembly

The estimated sugarcane genome is approximately 1Gbp. Our current non-redundant assembly is close to this size as well. It is worth remembering that current sugarcane cultivars are the result of interspecific hybridizations between polyploid parental species, and are themselves polyploid. Under that light, our current assembly is close to the monoploid genome size, but most likely it is collapsing the sequences of different homeologous chromosomes due to the shallow coverage. Nevertheless it should be a good resource to explore the sugarcane gene space, gene structures (exons, introns) and regulatory space (gene promoters).

In order to evaluate how representative is the current assembly in terms of gene space coverage we employed two software packages, CEGMA (PARRA *et al.* 2007) and BUSCO (SIMAO *et al.* 2015). Both of them look for sets of conserved single copy genes, the proportion of recovered genes gives a clear idea of the completeness of the assembly. CEGMA looks for a set of 248 conserved genes in eukaryotes, and it identified 77% of these genes in our assembly, in average every gene having approx. 3 copies in the assembly. Thus, this current assembly to some extent recovers some of the variants in the different homeologous chromosomes. BUSCO looks for 956 single copy genes conserved in the plant kingdom; it could identify approx. 70% in our assembly, with around half of these presenting more than one copy. These percentages can increase to 93% (CEGMA) and 79% (BUSCO) when considering partial hits, i.e., fragmented genes. These results clearly indicate that the current assembly have good representativity of the gene space, although

further sequencing would be required to have a complete view, and in order to resolver better the contributions of the different homologous chromosomes.

Identification of repeat sequences and Gene prediction

Table 4 shows a summary of the repeats identified by RepeatMasker. The most abundant sequence repeats are LTR retrotransposons of the types Copia and Gypsy, and DNA transposons. In total, interspersed repeats represent 45.95% of the total assembled sequence (Table 1).

For the identification of protein-coding genes we performed several rounds of genome annotation using the *ab initio* gene predictor Augustus (STANKE *et al.* 2006), BRAKER1 (HOFF *et al.* 2015) and PASA (HAAS *et al.* 2003). The first round of genome annotation only exploiting the genome sequence, i.e., *ab initio* (R1, see below) was run over the unmasked genome, additional rounds exploited RNASeq data, with the genome masked (R2, R6 – R8) or un-masked (R3 – R5, R9). Hits to the publicly available sugarcane ESTs at NCBI, and to the Sorghum proteins from Phytozome were computed with Exonerate. Different lines of evidence were combined using EVM (HAAS *et al.* 2008), to generate a high-quality genome annotation. Genome annotation is available, upon request (see Genome data availability below) through Web Apollo: http://bce.bioetanol.cnpem.br:8080/apollo. Table 5 lists the main results of the different protein-coding gene prediction rounds and some summary results. It is important to note here that exploiting information from the RNASeq data we can increase on the quality of predicted gene models.

Table 4. Frequency of repeat sequences in the draft assembly

Type of element	Number of elements	Length occupied	Percent of assembly
Retroelements	443,603	454,641,709 bp	38.86%
LINEs	43,710	25,695,257 bp	2.2%
L1/CIN4	35,062	22,527,373 bp	1.93%
LTR elemens	391,117	427,513,218 bp	36.54%
Ty1/Copia	106,650	119,109,620 bp	10.18%
Gypsy/DIRS1	280,341	306,930,668 bp	26.23%
DNA transposons	286,698	77,553,004 bp	6.63%
Tourist/Harbinger	94,473	19,946,309 bp	1.7%
Total interspersed repeats		537,765,472 bp	45.96%
Satellites	17,386	24,586,020 bp	2.1%

Table 5. Rounds of genome annotation and statistics.

A: Number of predicted protein-coding gene models; B: Percent of predicted gene models with CRBBs to Sorghum transcripts / Percent of sorghum transcripts with CRBBs to sugarcane gene models; C: Number of protein-coding transcripts with hits to PFAM / Number of PFAM models; D: BUSCO identified conserved genes, n=956, 'C'= Complete, 'D'= Duplicated, 'F'= Fragmented, 'M'= Missing

Round	Repeat status	Round	A	B	C	E
R1	*Off*	*ab initio* Augustus, using CEGMA genes for training	177,428	26 54	90,325 / 13,345	C:76%[D:47%] F:12% M:11%
R2	On	Annotation with BRAKER1, exploiting RNASeq data from (Mattiello et al. 2015).	197,834	26 58	48,678 / 12,356	C:80%[D:48%] F:10% M:8.6%
R3	*Off*	*ab initio* Augustus using the statistical model of R2.	589,866	12 57	73,346 / 11,523	C:42%[D:24%] F:29% M:28%
R4	*Off*	*ab initio* Augustus using the statistical model of R2 and using hints (intron positions derived from RNAseq data).	594,010	13 62	79,807 / 12,623	C:71%[D:37%] F:15% M:12%
R5	*Off*	Same as R4 but also predicting alternative spliced forms.	601,067	13 62	81,507 / 12,690	C:71%[D:39%] F:15% M:12%
R6	*On*	*ab initio* Augustus using the statistical model of R2.	335,714	21 56	42,264 / 10,501	C:40%[D:23%] F:30% M:29%
R7	*On*	*ab initio* Augustus using the statistical model of R2 and using hints (intron positions derived from RNAseq data).	341,380	22 61	48,887 / 11,863	C:71%[D:35%] F:15% M:13%
R8	*On*	Same as R7 but also predicting alternative spliced forms	347,505	22 61	50,389 / 11,952	C:71%[D:38%] F:15% M:12%
R9	*Off*	PASA Pipeline	113,906		67,939 / 17,953	C:77%[D:59%] F:8.3% M:13%

We have used EVidenceModeller (HAAS *et al.* 2008) to combine all the different sources of evidence listed in Table 5 into a high-quality set of predicted protein-coding genes, selecting a single gene model per locus. The set of transcripts or alignments used as evidence and their weights used in EVM are listed in Table 6. This resulted in a set of 153,078 predicted high-quality protein-coding genes, with 30.7% of them having hits to 23,433 PFAM domains. We compare the set of genes that do not appear to code proteins from the PASA

annotation set (R9) to the set of high-quality protein coding genes (EVM), and found a set of 29,267 putative Natural Antisense Transcripts, there are gene that map in opposite strands with an overlap of at least 100bp with identity of 99%.

Table 6. Weights for the difference types of evidence used to generate a high quality set of protein-coding genes using EVM

Evidence	Weight
Sugarcane EST alignments (Exonerate)	6
Sorghum protein alignments (Exonerate)	4
Ab initio prediction R2	5
Ab initio prediction R1	1
Ab initio prediction R3	1
Ab initio prediction R4	1
Ab initio prediction R6	1
Ab initio prediction R7	1
Ab initio prediction R9	10

Applying the approach to identify Transcription Associated Proteins (Transcription Factors and other Transcriptional Regulators), described in (PEREZ-RODRIGUEZ *et al.* 2010), to the set of high-quality protein-coding genes we could identify 3.815 TAPs belonging to 94 different protein families. Figure 4 shows the proteome size normalized distribution of TAPs in six angiosperms.

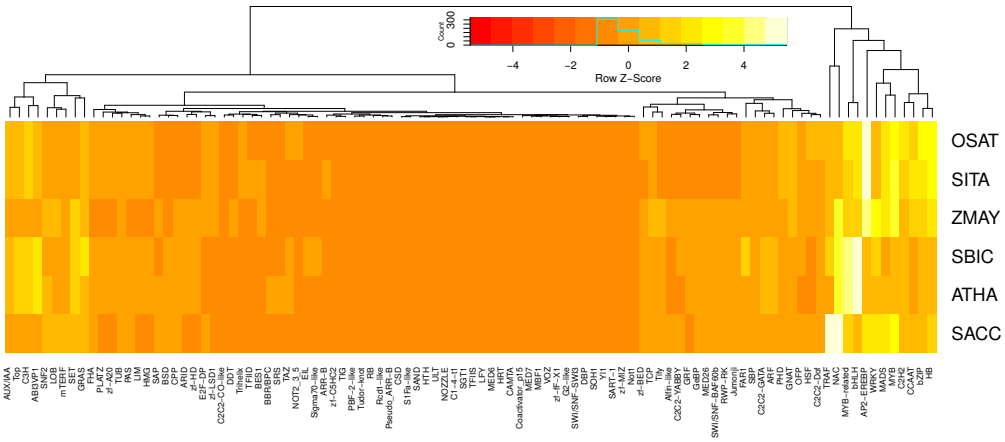


Figure 4. Identification and classfication of TAPs in siz Angiosperm species (ATHA: Arabidopsis thaliana, OSAT: Oriza sativa, SITA: Setaria itálica, SBIC: Sorghum bicolor, ZMAY: Zea mays: SACC: Saccharum SP80-3280)

Using the high quality prediction of protein-coding genes we aimed at identifying groups of orthologues genes between Sugarcane SP80-3280 (SACC) and the angiosperms *Arabidopsis thaliana* (ATHA), *Oryza sativa* (OSAT), *Setaria italica* (SITA), *Zea mays* (ZMAY), and *Sorghum bicolor* (SBIC). In order to achieve this we used the OrthoMCL pipeline with an inflation value of 1.5. Table 7 shows the total number of groups of orthologues found for each species, as well as the proportion of proteins that could be assigned to these groups. It is worth mentioning that there is a set of protein where the OrthoMCL pipeline can not

recognized enough similarity to any of the proteins in the whole data set, thus remaining as singletons, particularly for sugarcane this amount to 37.6% of all predicted proteins (Exclusive proteins in Table 1). Figure 5 shows shared groups of orthologous genes. There are 7800 groups of orthologous genes shared between all monocot and dicot species, suggesting the presence of 7800 ancestral functions in the most recent common ancestor of angiosperms. While all grasses share 3345 groups.

Table 7. Statistics about the clustering of orthologuous genes and exclusive proteins

Symbol	Predicted protein-coding genes	Groups of orthologues	Proteins in groups of orthologues	Exclusive proteins
SACC	153078	31,107	95,564 (62.4%)	57,514 (37.6%)
SBIC	39,441	20,338	33,500 (84.9%)	5,941 (15.1%)
ZMAY	88,760	23,919	56,781 (64.0%)	31,979 (36.0%)
SITA	40,599	18,720	31,566 (77.8%)	9,033 (22.2%)
OSAT	66.338	19.068	51.746 (78.0%)	14.592 (22.0%)
ATHA	35.386	12.160	30.704 (86.8%)	4.682 (13.2%)

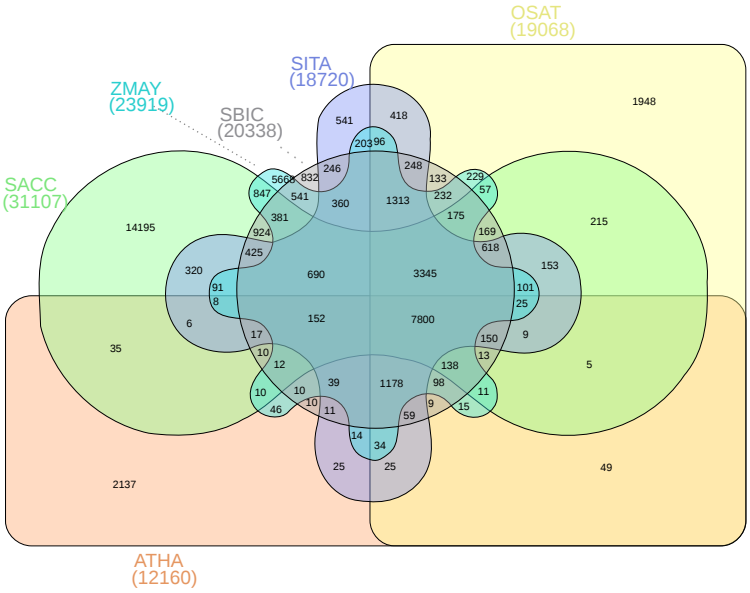
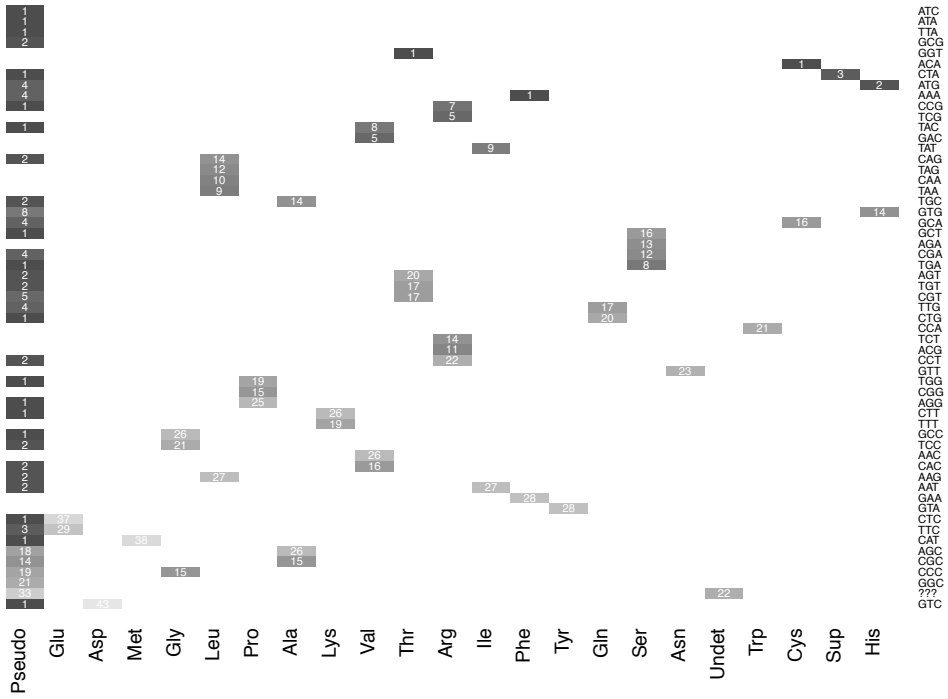


Figure 5. Groups of orthologues shared among the angiosperms Sugarcane SP80-3280 (SACC) Arabidopsis thaliana *(ATHA),* Oryza sativa *(OSAT),* Setaria italica *(SITA),* Zea mays *(ZMAY), and* Sorghum bicolor *(SBIC). Venn diagram generated with* http://www.interactivenn.net/

Aside from protein-coding genes, we identified 1,067 tRNA genes on 988 genomic contigs. 177 of these tRNAs are predicted pseudogenes, 3 are suppressor genes and 22 are undetermined. The remaining 865 tRNA carry anticodons for the standard 20 aminoacids (Figure 6). For contrast, the *Sorghum bicolor* genome has

649 tRNA, of these, 577 decode the standard 20 amino acids and 62 are predicted pseudogenes (http://lowelab.ucsc.edu/GtRNAdb/Sbico/).

Figure 6. Number of tRNA genes for the different aminoacids and anti-codons



Genome data availability

The genome raw data, assembly and annotation will be made available to Principal Investigators (Institutional e-mail required) after signing of the terms of the "Reserved analyses" notice shown in the appendix. In case of any doubts please contact diego.riano@bioetanol.org.br

Author contributions

DMRP and LM conceived the study. LM and LPC extracted and prepared the genomic DNA. All authors contributed to writing the final report. DMRP made all computational analysis and wrote the final version of the manuscript.

References

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS and D. J. LIPMAN, 1990 Basic local alignment search tool. J Mol Biol **215:** 403-410.

CARDOSO-SILVA, C. B., E. A. COSTA, M. C. MANCINI, T. W. BALSALOBRE, L. E. CANESIN *et al.*, 2014 De novo assembly and transcriptome analysis of contrasting sugarcane varieties. PLoS One **9:** e88462.

CARNAVALE BOTTINO, M., S. ROSARIO, C. GRATIVOL, F. THIEBAUT, C. A. ROJAS *et al.*, 2013 High-throughput sequencing of small RNA transcriptome reveals salt stress regulated microRNAs in sugarcane. PLoS One **8:** e59423.

CARUCCIO, N., 2011 Preparation of next-generation sequencing libraries using Nextera technology: simultaneous DNA fragmentation and adaptor tagging by in vitro transposition. Methods Mol Biol **733:** 241-255.

CHANDEL, A. K., S. S. DA SILVA, W. CARVALHO and O. V. SINGH, 2012 Sugarcane bagasse and leaves: foreseeable biomass of biofuel and bio-products. Journal of Chemical Technology & Biotechnology **87:** 11-20.

CHEAVEGATTI-GIANOTTO, A., H. M. DE ABREU, P. ARRUDA, J. C. BESPALHOK FILHO, W. L. BURNQUIST *et al.*, 2011 Sugarcane (Saccharum X officinarum): A Reference Study for the Regulation of Genetically Modified Cultivars in Brazil. Trop Plant Biol **4:** 62-89.

CHEVREUX, B., T. WETTER and S. SUHAI, 1999 Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. , pp. 45=56 in *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics* Heidelberg, Germany.

CORDEIRO, G. M., F. ELIOTT, C. L. MCINTYRE, R. E. CASU and R. J. HENRY, 2006 Characterisation of single nucleotide polymorphisms in sugarcane ESTs. Theor Appl Genet **113:** 331-343.

D'HONT, A., 2005 Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. Cytogenet Genome Res **109:** 27-33.

D'HONT, A., F. DENOEUD, J. M. AURY, F. C. BAURENS, F. CARREEL *et al.*, 2012 The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature **488:** 213-217.

DE SETTA, N., C. B. MONTEIRO-VITORELLO, C. J. METCALFE, G. M. CRUZ, L. E. DEL BEM *et al.*, 2014 Building the sugarcane genome for biotechnology and identifying evolutionary trends. BMC Genomics **15:** 540.

DUFRESNE, F., M. STIFT, R. VERGILINO and B. K. MABLE, 2014 Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. Mol Ecol **23:** 40-69.

FERREIRA, T. H., A. GENTILE, R. D. VILELA, G. G. COSTA, L. I. DIAS *et al.*, 2012 microRNAs associated with drought response in the bioenergy crop sugarcane (Saccharum spp.). PLoS One **7:** e46703.

FU, L., B. NIU, Z. ZHU, S. WU and W. LI, 2012 CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics **28:** 3150-3152.

GARCIA, A. A., M. MOLLINARI, T. G. MARCONI, O. R. SERANG, R. R. SILVA *et al.*, 2013 SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Sci Rep **3:** 3399.

GENTILE, A., T. H. FERREIRA, R. S. MATTOS, L. I. DIAS, A. A. HOSHINO *et al.*, 2013 Effects of drought on the microtranscriptome of field-grown sugarcane plants. Planta **237:** 783-798.

GRATIVOL, C., M. REGULSKI, M. BERTALAN, W. R. MCCOMBIE, F. R. DA SILVA *et al.*, 2014 Sugarcane genome sequencing by methylation filtration provides tools for genomic research in the genus Saccharum. Plant J **79:** 162-172.

GRIVET, L., and P. ARRUDA, 2002 Sugarcane genomics: depicting the complex genome of an important tropical crop. Curr Opin Plant Biol **5:** 122-127.

GUPTA, P., R. GOEL, S. PATHAK, A. SRIVASTAVA, S. P. SINGH *et al.*, 2013 De novo assembly, functional annotation and comparative analysis of Withania somnifera leaf and root transcriptomes to identify putative genes involved in the withanolides biosynthesis. PLoS One **8:** e62714.

HAAS, B. J., A. L. DELCHER, S. M. MOUNT, J. R. WORTMAN, R. K. SMITH, JR. *et al.*, 2003 Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res **31:** 5654-5666.

HAAS, B. J., S. L. SALZBERG, W. ZHU, M. PERTEA, J. E. ALLEN *et al.*, 2008 Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol **9:** R7.

HASSUANI, S., M., L. M and I. MACEDO, 2005 *Biomass power generation: sugar cane bagasse and trash*. PNUD-CTC, Piracicaba.

HOFF, K. J., S. LANGE, A. LOMSADZE, M. BORODOVSKY and M. STANKE, 2015 BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics.

HOTTA, C., C. LEMBKE, D. DOMINGUES, E. OCHOA, G. Q. CRUZ *et al.*, 2010 The Biotechnology Roadmap for Sugarcane Improvement. Tropical Plant Biology **3:** 75-87.

ILLUMINA, 2015 Long Read Sequencing Technology, pp.

JACKSON, S. A., A. IWATA, S. H. LEE, J. SCHMUTZ and R. SHOEMAKER, 2011 Sequencing crop genomes: approaches and applications. New Phytol **191:** 915-925.

KENT, W. J., 2002 BLAT--the BLAST-like alignment tool. Genome Res **12:** 656-664.

KOZOMARA, A., and S. GRIFFITHS-JONES, 2011 miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res **39:** D152-157.

LE CUNFF, L., O. GARSMEUR, L. M. RABOIN, J. PAUQUET, H. TELISMART *et al.*, 2008 Diploid/polyploid syntenic shuttle mapping and haplotype-specific chromosome walking toward a rust resistance gene (Bru1) in highly polyploid sugarcane (2n approximately 12x approximately 115). Genetics **180:** 649-660.

LI, F., G. FAN, K. WANG, F. SUN, Y. YUAN *et al.*, 2014 Genome sequence of the cultivated cotton Gossypium arboreum. Nat Genet **46:** 567-572.

LI, Z., Y. CHEN, D. MU, J. YUAN, Y. SHI *et al.*, 2012 Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Brief Funct Genomics **11:** 25-37.

LULIN, H., Y. XIAO, S. PEI, T. WEN and H. SHANGQIN, 2012 The first Illumina-based de novo transcriptome sequencing and analysis of safflower flowers. PLoS One **7:** e38653.

MAPA (Editor), 2012 *Anuário Estatístico da Agroenergia*.

MATTIELLO, L., D. M. RIAÑO-PACHÓN, M. C. M. MARTINS, L. P. DA CRUZ, D. BASSI *et al.*, 2015 Physiological and transcriptional analyses of developmental stages along sugarcane leaf, pp.

MCCOY, R. C., R. W. TAYLOR, T. A. BLAUWKAMP, J. L. KELLEY, M. KERTESZ *et al.*, 2014 Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PLoS One **9:** e106689.

MENOSSI, M., M. C. SILVA-FILHO, M. VINCENTZ, M. A. VAN-SLUYS and G. M. SOUZA, 2008 Sugarcane functional genomics: gene discovery for agronomic trait development. Int J Plant Genomics **2008:** 458732.

MYERS, E. W., and W. MILLER, 1988 Optimal alignments in linear space. Comput Appl Biosci **4:** 11-17.

MYERS, E. W., G. G. SUTTON, A. L. DELCHER, I. M. DEW, D. P. FASULO *et al.*, 2000 A whole-genome assembly of Drosophila. Science **287:** 2196-2204.

NEEDLEMAN, S. B., and C. D. WUNSCH, 1970 A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol **48:** 443-453.

NISHIYAMA, M. Y., JR., S. S. FERREIRA, P. Z. TANG, S. BECKER, A. PORTNER-TALIANA *et al.*, 2014 Full-length enriched cDNA libraries and ORFeome analysis of sugarcane hybrid and ancestor genotypes. PLoS One **9:** e107351.

OLSON, S. A., 2002 EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. Brief Bioinform **3:** 87-91.

ORTIZ-MOREA, F. A., R. VICENTINI, G. F. SILVA, E. M. SILVA, H. CARRER *et al.*, 2013 Global analysis of the sugarcane microtranscriptome reveals a unique composition of small RNAs associated with axillary bud outgrowth. J Exp Bot **64:** 2307-2320.

PARRA, G., K. BRADNAM and I. KORF, 2007 CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23:** 1061-1067.

PEREZ-RODRIGUEZ, P., D. M. RIANO-PACHON, L. G. CORREA, S. A. RENSING, B. KERSTEN *et al.*, 2010 PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Res **38:** D822-827.

R CORE TEAM, 2015 R: A Language and Environment for Statistical Computing, pp. R Foundation for Statistical Computing, Vienna, Austria.

ROACH, B. T., 1969 Cytological studies in Saccharum. Chromosome transmission in interspecific and intergeneric crosses. Proc. Int. Soc. Sugar Cane Technol. **13:** 901-920.

SIMAO, F. A., R. M. WATERHOUSE, P. IOANNIDIS, E. V. KRIVENTSEVA and E. M. ZDOBNOV, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31:** 3210-3212.

SOUZA, G., H. BERGES, S. BOCS, R. CASU, A. D'HONT *et al.*, 2011 The Sugarcane Genome Challenge: Strategies for Sequencing a Highly Complex Genome. Tropical Plant Biology **4:** 145-156.

STANKE, M., O. KELLER, I. GUNDUZ, A. HAYES, S. WAACK *et al.*, 2006 AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res **34:** W435-439.

VALDES, C., 2011 Brazil's Ethanol Industry: Looking Forward, pp. in *Report from the Economic Research Service*, edited by UNITED STATES DEPARTMENT OF AGRICULTURE.

VICENTINI, R., A. BOTTCHER, S. BRITO MDOS, A. B. DOS SANTOS, S. CRESTE *et al.*, 2015 Large-Scale Transcriptome Analysis of Two Sugarcane Genotypes Contrasting for Lignin Content. PLoS One **10:** e0134909.

XU, X., S. PAN, S. CHENG, B. ZHANG, D. MU *et al.*, 2011 Genome sequence and analysis of the tuber crop potato. Nature **475:** 189-195.

YANG, X., C.-Y. YE, Z.-M. CHENG, T. TSCHAPLINSKI, S. WULLSCHLEGER *et al.*, 2011 Genomic aspects of research involving polyploid plants. Plant Cell, Tissue and Organ Culture (PCTOC) **104:** 387-397.

YAZAWA, T., H. KAWAHIGASHI, T. MATSUMOTO and H. MIZUNO, 2013 Simultaneous transcriptome analysis of Sorghum and Bipolaris sorghicola by using RNA-seq in combination with de novo transcriptome assembly. PLoS One **8:** e62460.

ZANCA, A. S., R. VICENTINI, F. A. ORTIZ-MOREA, L. E. DEL BEM, M. J. DA SILVA *et al.*, 2010 Identification and expression analysis of microRNAs and targets in the biofuel crop sugarcane. BMC Plant Biol **10:** 260.

Appendix

Appendix A. “Reserved analyses” notice for access to the genome data generated at CTBE

As a public service, the Brazilian Bioethanol Science and Technology Laboratory – Brazilian Center for Research in Energy and Materials (CTBE/CNPEM) is making the draft sugarcane SP80-3280 genome sequence available before scientific publication.

By accessing these data, you agree not to publish any articles containing analyses of genes or genomic data on a whole genome or chromosome scale prior to publication by CTBE/CNPEM and/or its collaborators of a comprehensive genome analysis ("Reserved Analyses"). "Reserved analyses" include the identification of complete (whole genome) sets of genomic features such as genes, gene families, regulatory elements, repeat structures, GC content, or any other genome feature, and whole-genome - or chromosome-scale comparisons with other species.

The embargo on publication of Reserved Analyses by researchers outside of the Sugarcane SP80-3280 Draft Genome Sequencing Project at CTBE is expected to extend until the publication of the results of the sequencing project is accepted. Scientific users are free to publish papers dealing with specific genes or small sets of genes using the sequence data. If these data are used for publication, the following acknowledgment should be included: 'These sequence data were produced by the Brazilian Bioethanol Science and Technology Laboratory – Brazilian Center for Research in Energy and Materials'. If need come, this letter will be circulated to Journal Editors so that they are aware of the conditions of access and publication detailed above.

These data may be downloaded and used by all who respect the restrictions in the previous paragraphs, after signing this agreement. The assembly and sequence data should not be redistributed or repackaged without permission from the CTBE/CNPEM. Any redistribution of the data during the embargo period should carry this notice: "The Brazilian Bioethanol Science and Technology Laboratory – Brazilian Center for Research in Energy and Materials provides these data in good faith, but makes no warranty, expressed or implied, nor assumes any legal liability or responsibility for any purpose for which the data are used. Once the sequence is moved to unreserved status, the data will be freely available for any subsequent use."

We request that potential users of this sequence assembly contact us at diego.riano@bioetanol.org.br with their plans to ensure that proposed usage of sequence data are not Reserved Analyses.

To request access to the genome data, please e-mail this page, after filling in the fields below and signing it, to diego.riano@bioetanol.org.br. **Only institutional e-mail addresses from Principal Researchers will be allowed access.**

I _____, identified by (Doc type) _____ number _____, affiliated to: _____, located in address (institution): _____, city: _____, country: _____ hereby agree with the terms expressed in the “Reserved analyses” notice shown above.

_____

Signature

Telephone: (___) _____

Institutional E-mail: _____

Date:___/___/___, City:_____, Country: _____