GigaScience

Assembly of the 373K gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop

--Manuscript Draft--

Manuscript Number:	GIGA-D-19-00013R1
Full Title:	Assembly of the 373K gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop
Article Type:	Research

Funding Information:		
	FAPESP (2012/51062-3)	Professor Glaucia Mendes Souza
	FAPESP (2008/52146-0)	Professor Glaucia Mendes Souza
	FAPESP (2014/50921-8)	Professor Glaucia Mendes Souza
	FAPESP (2008/52074-0)	Not applicable
	FAPESP (2011/50761-2)	Not applicable
	National Science Foundation (DBI-1350041)	Not applicable
	CNPq (304360/2014-7)	Professor Glaucia Mendes Souza
	CNPq (308197/2010-0)	Not applicable
	FAPESP (2015/22993-7)	Not applicable
	FAPESP (2013/18322-4)	Not applicable
	FAPESP (2015/15346-5)	Not applicable
	CNPq (159094/2014-3)	Not applicable
	FAPESP (2017/02270-6)	Not applicable
	CAPES (DS-1454337)	Not applicable
	FAPESP (2013/23048-9)	Not applicable
	FAPESP (2016/06917-1)	Not applicable
	FAPESP (2013/07467-1)	Not applicable
	FAPESP (2017/02842-0)	Not applicable
	CNPq (309566/2015-0)	Not applicable
	National Science Foundation (IOS/0115903)	Not applicable
	National Institutes of Health (R01-HG006677)	Not applicable

Abstract:

Background
Sugarcane cultivars are polyploid interspecific hybrids of giant genomes, typically with 10-13 sets of chromosomes from two Saccharum species. The ploidy, hybridity and size of the genome, estimated to have in excess of 10 Gb, pose a great challenge for sequencing.
Results
Here we present a gene-space assembly of SP80-3280, including 373,869 putative genes and their potential regulatory regions. Their alignment to single copy genes of diploid grasses indicates that we could resolve 2-6 (up to 15) putative homo(eo)logs

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

	that are 99.1% identical within their coding sequences. Dissimilarities increase in their regulatory regions and gene promoter analysis shows differences in regulatory elements within gene families and are species-specific expressed. We exemplify these differences for sucrose synthase (SuSy) and phenylalanine ammonia-lyase (PAL), two gene families central to carbon partitioning. SP80-3280 have particular regulatory elements involved in sucrose synthesis not found in the ancestor S. spontaneum. PAL regulatory elements are found in co-expressed genes related to fiber synthesis within gene networks defined during plant growth and maturation. Comparison to sorghum reveals predominantly biallelic variations in sugarcane, consistent with the formation of two 'subgenomes' after their divergence ca. 3.8~4.6 MYA and reveals SNVs that may underlie their differences. Conclusions This gene-copy resolved assembly represents a large step towards a whole genome assembly of a commercial sugarcane cultivar providing a large diversity of genes and homo(eo)logs useful for improving biomass and food production.
Corresponding Author:	Glaucia Mendes Souza, Ph.D Universidade de São Paulo Sao Paulo, SP BRAZIL
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Universidade de São Paulo
Corresponding Author's Secondary Institution:	
First Author:	Glaucia Mendes Souza, Ph.D
First Author Secondary Information:	
Order of Authors:	Glaucia Mendes Souza, Ph.D
	Marie-Anne Van Sluys, Ph.D
	Carolina Gimiliani Lembke, Ph.D
	Hayan Lee, Ph.D
	Gabriel Rodrigues Alves Margarido, Ph.D
	Carlos Takeshi Hotta, Ph.D
	Jonas Weissmann Gaiarsa, Ph.D
	Augusto Lima Diniz, Ph.D
	Mauro de Medeiros Oliveira, Ph.D
	Sávio de Siqueira Ferreira, Ph.D
	Milton Yutaka Nishiyama-Jr, Ph.D
	Felipe ten Caten, Ph.D
	Geovani Tolfo Ragagnin, MSc
	Pablo de Morais Andrade, Ph.D
	Robson Francisco de Souza, Ph.D
	Gianlucca Gonçalves Nicastro, Ph.D
	Ravi Pandya, BS.c
	Changsoo Kim, Ph.D
	Hui Guo, Ph.D
	Alan Mitchell Durham, Ph.D
	Monalisa Sampaio Carneiro, Ph.D
	Jisen Zhang, Ph.D
	Qing Zhang, Ph.D

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

	Qing Zhang, Ph.D
	Ray Ming, Ph.D
	Michael Schatz, Ph.D
	Bob Davidson
	Andrew Paterson, Ph.D
	David Heckerman, Ph.D
Order of Authors Secondary Information:	
Response to Reviewers:	Dear Dr. Hans Zauner Firstly, we would like to thank the editor and reviewers for the valuable comments on our submitted manuscript. We have responded to all comments and criticisms and have revised the paper in light of them. A point-by-point response to these concerns is provided in the 'Response letter GIGA-D-19-00013.docx' file. We reinforce the high value of the presented sugarcane hybrid gene space assembly, not only for the sugarcane community, but also for those interested in unraveling genomics of complex polyploid crops. The revised version of our manuscript (in addition to Fig.2 and Additional files 1 and 2) has been uploaded.
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.	Yes

Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

Revised Manuscript
Click here to access/download;Manuscript;Revised GIGA-D-19-00013 Manuscript.docx
Click here to view linked References

Assembly of the 373K gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop

Full name	Institutional address	e-mail
Glaucia Mendes Souza*	1	glmsouza@iq.usp.br
Marie-Anne Van Sluys*	2	mavsluys@usp.br
Carolina Gimiliani Lembke	1	carolina.lembke@gmail.com
Hayan Lee	3,4	hayan.lee@stanford.edu
Gabriel Rodrigues Alves Margarido	5	gramarga@usp.br
Carlos Takeshi Hotta	1	hotta@iq.usp.br
Jonas Weissmann Gaiarsa	2	jonaswg@gmail.com
Augusto Lima Diniz	1	augustold@usp.br
Mauro de Medeiros Oliveira	1	mauromedeiros@usp.br
Sávio de Siqueira Ferreira	1,2	saviobqi@gmail.com
Milton Yutaka Nishiyama-Jr	1,6	yutakajr@gmail.com
Felipe ten Caten	1	ftencaten@gmail.com
Geovani Tolfo Ragagnin	2	geovaniragagnin@gmail.com
Pablo de Morais Andrade	1	pablo.andrade@gmail.com
Robson Francisco de Souza	7	rfsouza@usp.br
Gianlucca Gonçalves Nicastro	7	nicastro@iq.usp.br
Ravi Pandya	8	ravip@microsoft.com,
Changsoo Kim	9,10	changsookim@cnu.ac.kr
Hui Guo	9	huiguo7@gmail.com
Alan Mitchell Durham	11	aland@usp.br
Monalisa Sampaio Carneiro	12	monalisa@ufscar.br
Jisen Zhang	13	zjisen@126.com
Xingtan Zhang	13	tanger_009@163.com
Qing Zhang	13	zhangqing970@126.com
Ray Ming	13,14	rayming@illinois.edu
Michael C. Schatz	3,15	michael.schatz@gmail.com
Bob Davidson	8	bob.davidson@microsoft.com
Andrew Paterson	9	paterson@uga.edu
David Heckerman	8	heckerma@hotmail.com

1 – Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, Av. Prof. Lineu Prestes, 748, São Paulo, SP 05508-000, Brazil

2 – Departamento de Botânica, Instituto de Biociências, Universidade de São Paulo, Rua do Matão, 277, São Paulo, SP 05508-090, Brazil

3 – Cold Spring Harbor Laboratory, One Bungtown Road, Koch Building #1119, Cold Spring Harbor, NY 11724, United States of America

4 – Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA CA 94598, United States of America

5 – Departamento de Genética, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Avenida Pádua Dias, 11, Piracicaba, SP 13418-900, Brazil

6 – Laboratório Especial de Toxinologia Aplicada, Instituto Butantan, Av. Vital Brasil, 1500, São Paulo, SP 05503-900, Brazil

7 – Departamento de Microbiologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, Av. Professor Lineu Prestes, 1734, São Paulo, SP 05508-900, Brazil

8 – Microsoft Research, One Microsoft Way, Redmond, WA 98052, United States of America9 – Plant Genome Mapping Laboratory, University of Georgia, 120 Green Street, Athens, GA 30602-7223, United States of America

10 – Department of Crop Science, Chungnam National University, 99 Daehak Ro Yuseong Gu, Deajeon, 34134, South Korea

11 – Departamento de Ciências da Computação, Instituto de Matemática e Estatística, Universidade de São Paulo, Rua do Matão, 1010, São Paulo, SP 05508-090, Brazil

12 - Departamento de Biotecnologia e Produção Vegetal e Animal, Centro de Ciências Agrárias, Universidade Federal de São Carlos, Rodovia Washington Luis km 235, Araras, SP 13.565-905, Brazil

13 – FAFU and UIUC-SIB Joint Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Shangxiadian Road, Fuzhou 350002, Fujian, China

14 - Department of Plant Biology, University of Illinois at Urbana-Champaign, 201 W. Gregory Dr. Urbana, Urbana, Illinois 61801, USA

15 – Departments of Computer Science and Biology, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218-2608, United States of America

*These authors contributed equally to this work and are co-corresponding authors: glmsouza@iq.usp.br and mavsluys@usp.br

ABSTRACT

Background

Sugarcane cultivars are polyploid interspecific hybrids of giant genomes, typically with 10-13 sets of chromosomes from two Saccharum species. The ploidy, hybridity and size of the genome, estimated to have in excess of 10 Gb, pose a great challenge for sequencing.

Results

Here we present a gene-space assembly of SP80-3280, including 373,869 putative genes and their potential regulatory regions. Their alignment to single copy genes of diploid grasses indicates that we could resolve 2-6 (up to 15) putative homo(eo)logs that are 99.1% identical within their coding sequences. Dissimilarities increase in their regulatory regions and gene promoter analysis shows differences in regulatory elements within gene families and are species-specific expressed. We exemplify these differences for sucrose synthase (SuSy) and phenylalanine ammonia-lyase (PAL), two gene families central to carbon partitioning. SP80-3280 have particular regulatory elements involved in sucrose synthesis not found in the ancestor *S. spontaneum*. PAL regulatory elements are found in co-expressed genes related to fiber synthesis within gene networks defined during plant growth and maturation. Comparison to sorghum reveals predominantly biallelic variations in sugarcane, consistent with the formation of two 'subgenomes' after their divergence ca. 3.8~4.6 MYA and reveals SNVs that may underlie their differences.

Conclusions

This gene-copy resolved assembly represents a large step towards a whole genome assembly of a commercial sugarcane cultivar providing a large diversity of genes and homo(eo)logs useful for improving biomass and food production.

Keywords: Allele; Bioenergy; Biomass; Genome; Polyploid

BACKGROUND

Sugarcane is the world's most cultivated crop in tonnage (more than rice, maize and wheat) [1], and is considered the most sustainable of energy crops [2] with high potential to mitigate climate change without affecting food security [3]. Already produced in over 100 countries, high productivity of sugar, bioethanol and bioelectricity [4] make it a highly expandable green alternative to petroleum [5–7]. The International Energy Agency projects a 150 EJ (17% of energy demand) contribution of bioenergy by 2060, delivering 18% of the emission reductions needed to achieve the 2DS (2ºC Scenario). Sugarcane bioenergy production by 2045 could displace up to 13.7% of crude oil consumption and 5.6% of the world's CO_2 emissions relative to 2014. This can be achieved without using forest preservation areas or land necessary for food production systems. Additionally, the myriad of products that can derive from sugarcane biomass [8] further enhance opportunities for sugarcane in a portfolio of technologies needed to transition to a low carbon 'bioeconomy'.

Opportunities to accelerate breeding progress and enrich knowledge of the fundamental biology of this important plant motivate efforts to produce a high-quality reference genome, a challenge that is unusually complex. Unlike wheat cultivated species known to be either tetraploid (AABB) or hexaploid (AABBDD), the *Saccharum* (sugarcane) genus is considered to be a species complex. A recent study [9] proposed independent polyploidization events within *Saccharum* after divergence from the last ancestor shared with *Sorghum*, superimposed upon an additional whole genome duplication since the diversification of grasses. As a consequence, the sugarcane genome is redundant and harbors genes in multiple functional copies. Adding further complexity, sugarcane cultivars are polyploid interspecific hybrids, typically with 10-13 sets of their 10 basic chromosomes, 80-85% from *Saccharum officinarum* (2n=80), which is known for its sweetness, 10-15% from *S. spontaneum* (2n=40-128) known for its robustness, and ~5% with recombined chromosomes between those two progenitors [10,11]. The ploidy, hybridity and sheer size of the genome, estimated to have in excess of 10 Gb, pose a great challenge for sequencing [12]. Recently released sequences of the modern cultivar R570 yielded a mosaic monoploid reference (382 Mb single tiling path) [13] and a *S. spontaneum* AP85-441 haploid assembly (3.13 Gb) [14].

Worldwide sugarcane yield (~84 ton/ha) is currently only ~20% of the theoretical potential (~381 ton/ha), spurring great interest in conventional or molecular breeding approaches to improve it. However, progress by conventional breeding towards closing the gap between current and potential yield has been slow with gains in the order of 1.0–1.5% a year [15]. Sugarcane commercial cultivars distribute roughly one third of their

carbon into sucrose and two thirds into tops and stems which, due to high lignin content, are burned to fuel boilers, contributing to the favorable energy balance of industrial processes [16]. As sugarcane can accumulate large amounts of sucrose in its stems, up to ~650 mM [17], it is important to study sucrose metabolism and the key players in its regulation. Also, of interest is the revealing of regulators of cell wall biosynthesis. Altering these pathways may help shift carbon partitioning from sucrose storage to biomass accumulation, rich in fiber content, mostly composed of secondary cell walls formed by cellulose, hemicellulose and lignin [18]. The latter compound is a hydrophobic polymer that provides strength and rigidity to the plant, but also is responsible for cell wall recalcitrance, which is the natural plant resistance to hydrolytic attacks that hampers cellulosic ethanol production [19].

RESULTS

The SP80-3280 assembly reveals a gene space of 373,869 genes

Here, we report a representative gene space assembly of the genome sequence of SP80-3280 (GenBank accession number QPEU01000000), the cultivar used in Brazilian breeding programs with the largest collection of transcriptomic data available [20]. On average, 6 sugarcane haplotypes, putatively homo(eo)logs, could be resolved from 4.26 Gb of assembled data and 373,869 putative genes and promoter regions. This is the first release of a resolved gene assembly of such a giant hybrid polyploid genome and their potential regulatory regions.

The assembly was constructed using 26 libraries sequenced using Illumina Synthetic Long-Read technology, obtaining 19 Gb, ~19x haploid genome coverage (~1.9X genome coverage) with >99% of bases having >99% accuracy (**Additional file 1: Fig. S1**). The final assembly includes 450,609 contigs (unitigs + singletons), with average length of 9,452 bp and NG50 of 41,394 bp (**Table 1)**, adding over 3Gb of sequence not previously reported (**Additional file 2: Table S1**) [21]. The gene space described here might be explored through a GBrowse environment available at http://sucest-fun.org/cgi-bin/cane_regnet/gbrowse2/gbrowse/microsoft_genome_moleculo_scga7/ or http://sucest-fun.org/wsapp/ (> Cane Genome > Sugarcane Genome Microsoft-Moleculo).

Several indicators support the comprehensiveness of the SP80-3280 gene space: (*i*) among 39,441 sorghum transcripts, 39,207 (99.4%) matched the assembly, at least partially; of these, 71.1% matched at least one sugarcane contig with 90% or higher coverage (**Additional file 1: Fig. S2**); (*ii*) the assembly completely covers 217 (87.5%) of the 248 ultra-conserved Core Eukaryotic Genes Mapping Approach (CEGMA) [22] proteins, and partly covers 18 (7.3%), with only 13 (5.2%) not detected (**Additional file 2: Table S2)**; (*iii*) among 1,440 genes in the Benchmarking Universal Single-Copy Orthologs (BUSCO) [23] Plantae lineage, the assembly completely covers 1,309 (90.9%) and partially covers 53 (3.7%) (**Additional file 2:Table S3)**. By including tBLASTn of the 78 (5.4%) missing Plantae lineage BUSCO genes, only 8 (0.5%) are absent; (*iv*) assembled chloroplast (NC_005878.2) and mitochondrial (LC107874.1and LC107875.1) genomes were over 99% similar (at gene level) to published *Saccharum* genomes [24,25]; and (*v*) 94.9% of 134,840 SP80-3280 expressed sequence tags (ESTs) match the assembled gene-space sequence.

The assembly revealed 373,869 putative genes with 374,774 transcripts (**Table 1**), far more than the 72,269 unigenes inferred from six sugarcane genotypes [26]; 85,151 transcripts of sugarcane genotypes with contrasting lignin contents [27]; and 195,765 transcripts inferred from *de novo* assembly of ORFeomes from *S. officinarum*, *S. spontaneum* and SP80-3280 [28]. The number of genes, high quality of alignments, and the following analysis indicates that the assembly provides a high-quality resolution of homo(eo)logous genes.

Among the predicted transcripts, 302,627 (80.7%) aligned to a Uniref50 protein [29], and 195,651 were annotated with 10,362 GO terms [30] (**Additional file 1: Fig. S3**). Our previously published SP80-3280 ORFeome was reassembled using the genome as a reference, revealing 269,050 genes and 275,807 transcripts from leaves, immature and intermediate internodes (**Additional file 2: Table S4)**. Further, a set of 134,840 SP80-3280 ESTs from a Sugarcane EST Project – SUCEST [20] – were mapped to assembled contigs and compared to predicted genes, in order to further estimate the completeness of the predicted gene space. A total of 127,940 ESTs (92.8%) have at least one match in the assembly, which is in accordance with similar analysis of other plant genomes [31], and only 6.8% of aligned ESTs (8,499) do not correspond with predicted genes. This result resembles the BUSCO results, for which only 5.4% of conserved genes could not be identified in the assembly. Although 10.4% of ESTs (12,966) have a unique hit, what may represent sequencing/assembly issues or genes loss, 84.9% of ESTs (106,133) show 2 up to 15 matches on the genome, reflecting the presence of the majority of putative homo(eo)logs (**Additional file 1: Fig. S4A**). This result is similar to the search of

CEGMA matches against the genome itself using BLASTn. From 235 sequences completely or partially covering CEGMA proteins, 205 has 2 or more hits on the assembly, with most of them (32) with 5 matches (**Additional file 1: Fig. S4B**).

To verify how the assembled genes reflected the expected content of homo(eo)logous genes, the gene content was compared to those of other grasses. Single-copy genes from diploid grasses (sorghum, rice and *Brachypodium*) are present in up to 15 copies in sugarcane, mostly with 2-6 copies (total of 1,592 coding sequences (CDS) in sugarcane) (**Fig. 1A**). Dissimilarities among putative homo(eo)logs increase from the coding region to the promoter region, with median divergence of 0.90% between CDS, 1.03% for the 100 nucleotides (nt) upstream, 4.47% for 500 nt and 7.50% for 1,000 nt (**Fig. 1B**). Frame-preserving INDELs are more abundant than frameshifts (**Fig. 1C**) and short frameshift INDELS were relatively less frequent in the sugarcane exons than in sorghum [32].

The SP80-3280 gene series that correspond to diploid grass single-copy genes showed expression of sense copies for multiple homo(eo)logs (**Fig. 2A**), with very few copies transcribed in antisense orientation (**Fig. 2B**) based on alignment with the SP80-3280 cDNA reads [28] from leaves, immature and intermediate internodes. For some genes, not all copies are expressed in SP80-3280 (**Fig. 2A, Additional file 1: Fig. S5 A**). In addition, the increase in the number of expressed copies is not accompanied by an increase in the level of expression (**Additional file 1: Fig. S5B**).

As an example of the complexities in data mining of such an intricate gene-space for future reference, we offer an example using two well-known genes involved in sucrose and lignin biosynthesis.

Gene family analysis of SuSy and PAL shows differences in their regulatory regions in SP80-3280 and *S. spontaneum*

Sucrose Synthases (SuSy) catalyze the reversible breakdown of sucrose into UDP-glucose and fructose in carbon partitioning [33]. In agreement with previous work on sugarcane progenitors [34] (*S. officinarum*, *S. robustum* and *S. spontaneum*), phylogenetic analysis of 44 ScSuSy (Sugarcane Sucrose Synthase) CDSs identified in the SP80-3280 assembly supports that this hybrid has 5 SuSy genes (hereafter ScSuSy1-5) in three groups: I (ScSuSy1 and 2), II (ScSuSy3 and 5) and III (ScSuSy4) (**Fig. 3A)**. Sorghum shares these 5 SuSy genes, indicating that they evolved before the sugarcane/sorghum divergence. RNA-Seq data from leaves

and internodes of SP80-3280 [28] shows expression of 36 of the 44 ScSuSy members, suggesting ScSuSy1-2 (group I) and ScSuSy5 might control carbon flux from source to biomass conversion in stems, as they show higher expression in internodes than in leaves (**Fig. 3C**).

Different members of the SuSy gene family may have different functional roles and in sugarcane this was observed as different expression levels related to different TFBs identified. We identified five different top-ranked TFBs (with the highest score) in the ScSuSy1-5 members. Three of them are related to auxin and abscisic-acid hormone signaling (ScSuSy1, 3, 5). For ScSuSy1 genes, the TFBS analysis predicted the motif wATATATATw (MA1184.1) that is associated with RVE1, a morning-phased transcription factor integrating the circadian clock and auxin pathway genes that bind to the evening element (EE) of promoters [35]. For ScSuSy2 genes, we found the motif GACrAATryA (MA1374.1) that is associated with IDD which regulates photoperiodic flowering by modulating sugar transport and metabolism [36]. For ScSuSy3 genes, we found the AyACTAGTrT (MA0930.1) motif in 64% of its SP80-3280 copies and in all copies in the *S. spontaneum* and R570 monoploid genomes. It is associated with ABA-responsive elements (ABRE) that regulate stress response via ABA signaling. For ScSuSy4 genes, we found the TAGyAynTTT (MA1012.1) motif that is probably involved in regulation of the photoperiod and vernalization pathways. Finally, for ScSuSy5 genes, we found a CTGCTAGCAG (MA0564.1) conserved motif exclusively for ScSuSy5 genes in SP80-3280. This motif allows binding with an element associated with ABI3, which participates in abscisic acid (ABA)-regulated gene expression. Previous studies from our group had already pointed out ABA- and sucrose-induced genes associated with higher sucrose content in sugarcane [37].

SuSy produces the substrate for cellulose biosynthesis (UDP-glucose) and is commonly associated with cell wall and cellulose synthesis [38,39]. In view of the myriad of possibilities to convert lignocellulosic compounds into chemicals and fuels, defining phenylpropanoid biosynthesis pathway members in sugarcane is of great interest. Phenylalanine ammonia-lyase (PAL) is the first enzyme in phenylpropanoid biosynthesis [40–42] and silencing its expression has been associated to a reduction in lignin content [40–43]. Lignin is a major component of plant cell walls [18], and is responsive to the ethylene-releasing ripener (ethephon) in both leaf and internode [44].

Mapping of predicted proteins from SP80-3280 against the SUCEST-FUN Cell Wall Catalogue [43] (731 transcripts of 20 protein categories) identified 3,054 similar proteins (**Additional file 2:Table S5)**, including

47 PAL copies. Phylogenetic analysis together with sorghum, *S. spontaneum* and mosaic monoploid R570 PAL sequences reveals 5 clusters (**Fig. 3B**), each containing at least one representative with a sorghum ortholog. *S. spontaneum* has 33 putative PAL genes, somewhat more than expected considering that the sequenced genotype is a tetraploid. The higher number may be due to expansion of PAL members in group I that occurred also for sorghum and the sugarcane hybrid genomes of R570 and SP80-3280. Group V has a higher number of SP80-3280 PAL members and all except one (ID 37780.4) showed expression evidence (**Fig. 3D**).

Regarding TFBS prediction within PAL regulatory sequences, we identified four different top-ranked TFBS. For PAL I, it was predicted an ArCAyATnTG (MA0930.1) element, which is associated with ABF3, a transcription factor involved in ABA and stress responses and acting as a positive component of glucose signal transduction. For PAL III, we found the element GGTCsGGCkC (MA0992.1), an element associated with AP2/ERF, a transcription factor involved in the regulation of gene expression by stress factors and by components of stress signal transduction pathways. For PAL Va, we found the element TCTAAAGTTT (MA0064.1), which is associated with PBF, a transcription factor involved in ABA, stress response and components of stress signal transduction pathways. Finally, for PAL Vb, we found the motif GCCGGAACGG (MA1009.1). This element is associated with ARF3, a transcription factor involved in auxin and ABA-regulated gene expression. In summary, our results corroborates reported findings [37] which reveal that PAL genes were induced by ABA.

In addition to PAL members expansion in group I, the CCR (Cinnamoyl-CoA reductase), COMT (Caffeic acid 3-O-methyltransferase) and 4CL (4-coumarate-CoA ligase) gene families, also related to phenylpropanoid biosynthesis, have much higher numbers of genes (620, 453 and 375, respectively) in sugarcane than sorghum [45] (44, 41 and 15, respectively). This is another challenge and opportunity for future functional characterization (**Additional file 2: Table S6)**.

The sheer number of sugarcane genes found so far, the large size of multi-gene families and the evidence that not all homo(eo)logs are expressed point to a very complex role of regulation in the determination of phenotypic differences. Consistent with the gene copy-richness of sugarcane, we inferred 15,737 transcription factors (TFs) from 57 families (**Additional file 2: Table S7)**, versus ~2,000 previously estimated [46]. The classification of core promoters and identification of Transcription Factor Binding Sites (TFBSs) in proximal

promoters was performed *in silico* and the percentage of core promoter regions with a TATA-box element was 47.72% and 12.76% for SuSy and PAL genes, respectively.

The TFBS identification pointed to a wealth of regulatory elements differentially distributed among members of the same gene family, i.e. SuSy and PAL (**Fig. 3C and D and Additional file 2: Table S8)**. In addition, using gene expression data of SP80-3280 plants grown in field conditions for 13 months, we have found evidence of a co-expression module, enriched for phenylpropanoid and lignin biosynthesis gene ontology terms (**Additional file 1: Fig. S6A**). This module comprises 116 transcripts, including one PAL (**Additional file 1: Fig. S6B**), whose expression is higher in internodes 5 and 9, than in leaves and immature internode (**Additional file 1: Fig. S6C**). It was possible to identify the TFBSs, predicted as putative regulators of the PAL gene family (**Fig. 3D**) within the upstream region of these co-expressed genes, suggesting that ABF, ERF, ZF-HD/C2H2, and ARF3 (**Additional file 1: Fig. S6D**) may also regulate other genes involved in lignin biosynthesis and metabolism. The most significant motifs found for each gene family (SuSy and PAL) were mapped to the promoter region of the remaining sequences from both SP80-3280 and R570 hybrids and *S. spontaneum* (**Additional file 2: Table S8 and Table S9**). Interestingly, only ScSuSy2 and ScSuSy3 motifs mapped in all species, suggesting that SP80-3280 hold particular regulatory elements involved in sucrose synthesis. Conversely, SP80-3280 and *S. spontaneum* share all predicted motifs for PAL genes (**Additional file 2: Table S9)**, suggesting that this gene family may be derived from the *S. spontaneum* ancestor.

Transposable element insertions may affect SuSy and PAL expression

Fewer transposable elements (TE) were identified in SP80-3280 gene space than in the AP85-441 *S. spontaneum* and mosaic monoploid R570 assembly, probably due to repetitive regions collapsing in the assembly even with the use of long synthetic-read sequencing (**Additional file 1: Fig. S7**, **Additional file 2: Table S10)**. All previously described TE families are represented in the three genome assemblies, disclosing few cultivar specific amplifications. The two modern cultivars (SP80-3280 and R570) have fewer TE counts than the *S. spontaneum* progenitor in normalized monoploid genomes. LTR retrotransposons are large contributors to genome composition at the chromosome assembly level. However, scMaximus (Copia) and scDel (Gypsy) LTR-retrotransposon families are similarly represented in both gene-space and chromosome assemblies supporting their presence in transcriptionally active regions [47]. We also note that scCACTA

transposons are more represented at the gene-space assembly than schAT while the scMutator family is similarly represented in both.

Functionally important TE insertions were identified in the ScSuSy gene family (**Fig. 3**). ScSuSy2 copies have a contrasting pattern, most *S. spontaneum* having TE insertions while most SP80-3280 homo(eo)logs do not – although SP80-3280 and *S. spontaneum* share one ancient insertion of schAT159 at similar distances from the ATG. ScSuSy3 genes are polymorphic between species and within SP80-3280, with 6 copies having no TE and 5 in which different TEs may impact expression. In particular, scga7_uti_cns_0020964:7575-17575 (-) harbors a full LTR at 280 bases from the ATG. Most ScSuSy4 copies have no TE insertion but interestingly, as described for ScSuSy2, SP80-3280 (scga7_uti_cns_0226458:7638-16073 (-)) and *S. spontaneum* (Chr1B:33406669-33416669 (-)) share one ancient schAT159 insertion. Finally, ScSuSy1 has similar patterns of TE presence and absence in both genomes, and ScSuSy5 genes have no insertions in the promoter regions of either *S. spontaneum* or SP80-3280. Furthermore, PAL genes from group I exhibit most of the copy variation and harbor TEs inserted near the promoter region. Only two copies from SP80-3280 and *S. spontaneum* lack TE insertion in PALs from group I.

Sugarcane and sorghum polymorphisms support recent allotetraploidy and suggest candidate genes for morphological and physiological differences between these taxa

Despite a common foundation for evolving high sugar content with similar SuSy genes (ScSuSy1-5), sugarcane and closely related sorghum have taken different paths since sharing ancestry. We identified 10,586 natural SNP variations (SNVs) between sorghum and sugarcane 4,140 unique genes, mostly bi-allelic (80.8%), but 6.2% tri-allelic and 0.97% tetra-allelic (**Fig. 4**). The overwhelming predominance of biallelic variations indicates that many sorghum genes are represented by two discernible sugarcane copies, supporting the theory of allotetraploidization shortly after divergence with sorghum ca. 3.8~4.6 MYA [48], creating two sugarcane 'subgenomes'. Recently published results from Vieira et al. [49], demonstrate that sugarcane meiotic chromosomes behave as bivalents, supporting this inference. Autotetraploidization after Saccharum speciation ca. 3.1~3.8 MYA may have further contributed to allelic richness within each sugarcane 'subgenome'. The preservation of as many as four functionally different alleles at a locus, with cases observed on all except one chromosome (Chr 10 - **Fig. 4**), is consistent with the well-known heterozygosity of sugarcane cultivars and

associated susceptibility to inbreeding depression. However, genes for which sugarcane has only one allele are more abundant than 3- or 4-allele loci, perhaps reflecting cases in which a single gene copy is sufficient, or in which occasional exchanges between subgenomes have homogenized multiple homo(eo)logs.

Further, 1,334 SNVs that differentiate sugarcane from sorghum in 585 single copy genes include frameshifts, premature termination, erroneous splicing, loss of stop codons and incorrect translation initiation (**Additional file 1: Fig. S8**, **Additional file 2: Table S11)** in genes significantly enriched in transcription, DNA-dependent cell organization and biogenesis in the nucleus and endoplasmic reticulum (**Additional file 2: Table S12)** comprise a rich slate of candidates for causes of morphological and physiological differences between these taxa.

The gene-space contribution towards a chromosome level assembly of a sugarcane commercial hybrid

Notwithstanding the fragmented nature of our assembly, we explored how it could contribute beyond the gene space toward a whole genome assembly of the hybrid sugarcane genome. Previous analysis of grass genomes revealed extensive conservation of gene order overlaid with a background of small-scale chromosomal rearrangements and numerous localized gene deletions, insertions and duplications [50]. Recently published estimates of the levels of gene synteny between *Sorghum bicolor* and the sugarcane cultivar R570 found that 83% of the genes are arranged co-linearly in the two genomes [13]. In our assembly of SP80-3280, 79,094 (17.6%) contigs had at least two predicted genes and could therefore be used to compare the order of genes in SP80-3280 to those of sorghum. To avoid the need to resolve multiple comparisons to duplicated regions in the sorghum genome, we generated a sequence similarity-based clustering of all coding sequences from both genomes and used the genes in clusters with only one sorghum gene as anchors to evaluate synteny (**Additional file 1: Fig. S9**). We found that 9,319 (2.1%) SP80-3280 contigs had at least two synteny anchors and 85% (7,906 – 1.8% of all contigs) of these contigs were fully syntenic (**Additional file 1: Fig. S10A, B**), *i.e.* had all genes in the same order and orientation in SP80-3280 contigs and the sorghum chromosomes (**Additional file 2: Table S13)**. To evaluate the effect of SP80-3280 assembly fragmentation on the number of segments with conserved gene order ("syntenic blocks") per contig, we used a Monte Carlo method to simulate the fragmentation of the chromosomes and contigs of the *Saccharum* R570 and *S. spontaneum* genomes. We performed 1,000 rounds of simulation for each genome and, at each round, sampled 10,000 random fragments from each of these two genomes, while simultaneously sampling the same number

of contigs from SP80-3280's assembly. Sampled contigs and contig fragments were constrained to follow the distribution of the number of genes per contig observed for the full SP80-3280 assembly. The number of syntenic blocks on each fragment was then evaluated and the relative frequency of contigs/fragments per number of syntenic blocks is shown in additional file 1, Fig. S10C. We observed that contigs and fragments harboring a single syntenic block are sampled at similar frequencies in all genomes analyzed. While an increase in sequencing coverage would lead to improved estimates of co-linearity, our analysis of the small subset of contigs with two or more marker genes suggests that levels of genomic rearrangement in SP80-3280 are similar to those expected anywhere in the genomes of the other two *Saccharum* species.

Finally, to allocate the gene space into potential physical groupings we aligned the SP80-3280 transposable element (TE) masked BWA-SW to chromosome level assemblies of the *S. spontaneum* tetraploid AP85-441 genome [14] and the R570 [13] monoploid genome data. Multiple correspondence analysis (MCA) with hierarchical clustering of the sequences enabled us to allocate the gene space contigs into 6 clusters, an important contribution to future scaffolding efforts. From the total of 450,609 contig sequences, 418,471 (92,86%) produced a BWA-SW alignment against the *S. spontaneum* [14] and R570 [13] assemblies (**Fig. 5A**) and protein alignment among these three species are consistent with MCA results (**Fig. 5B and C**). Contigs were also mapped against a collection of 778 targeted sequenced BACs of which 347 are from SP80-3280 and 431 from R570. All BACs had a corresponding contig match against the assembly. This collection shows centromeric regions and non-TE multigene families are the most covered (64x). An R gene locus (I2C-2) found in cluster 3 of SP80-3280 and in chromosome 9 of R570, was verified for co-location with a Ca^{+}-dependent kinase, a *dog1* (delay of germination 1) and an aminotransferase. The co-location was confirmed in R570 and SP80-3280 BACs showing up to eight copies of each gene (**Additional file 1: Fig. S11**).

DISCUSSION

This assembly presents 373,869 genes. The gene space described here represents a significant step in understanding the haplotype origin of the hybrid genome. Approximately 12.25% of the SP80-3280 genome sequence is of *S. spontaneum* origin [14], supporting previous studies [10,11]. The comparison against different sets of genes (sorghum, CEGMA, BUSCO, mitochondrial and chloroplast) supported the comprehensiveness of the gene space. The total of predicted genes (373,869) is around 10x, 14x and 13x higher

than those for monoploid genome assemblies of *S. spontaneum* [14], sugarcane R570 [13] and sorghum [52], respectively. This is in agreement with the predicted 8 to 14 copies for *S. spontaneum*, depending on the cytotypes, and for modern sugarcane varieties [53].

Although for sugarcane modern variates we expect eight or more copies of each chromosome, it is possible that each homolog does not contains a copy of every gene, because of potential gene loss. In addition, it is also possible that some homeologs were not identified in our assembly because of assembly or sequencing difficulties in regions with highly repetitive sequences. Thus, for sugarcane genes found to correspond to single-copy genes in diploid grasses, or that matched to CEGMA genes or that had an EST correspondent, we found mostly from 2 to 15 copies in the SP80-3280 assembly and the sequence differences are present mainly in the upstream regulatory region. This highlights the importance and complexity of studying homo(eo)logs expression in sugarcane and adds great value to the development of molecular markers for breeding in gene promoter regions. The differences in gene upstream sequences may potentially affect the expression level among the copies and across the studied tissues. This was also reported for the polyploids cotton [54] and wheat [55]. Expression differences among homo(eo)logs in polyploid species may play a crucial role in increasing adaptability to environmental stresses (such as salinity [56], heat and drought [57]) and in improving performance of new cultivars. These differences highlight the importance of our assembly which discriminates homo(eo)logs, for example providing important information for the selection of target sequences (genes or promoters) to produce transgenic sugarcane plants. With the homo(eo)logs identified, one could discard a sequence that is not expressed or use genome editing tools to modify a target sequence to increase its expression. It is also possible to identify the progenitor contributing a homo(eo)log (e.g., *S. spontaneum*, *S. officinarum* or a parent in a cross) and select the homo(eo)log from the progenitor that has the phenotype of interest.

In an attempt to organize the contigs, we allocate them in 6 clusters using MCA with hierarchical clustering of the sequences. The majority of proteins predicted from chromosomes 1, 2, 3 and 4 (in both *S. spontaneum* and R570) have their best matches located in SP80-3280 contigs from clusters 2, 5, 6 and 1, respectively (Fig. 5B and C). On the other hand, clusters 3 and 4, which contain contigs matching to multiple chromosomes, including those in which chromosomal rearrangement events were demonstrated in comparison

to sorghum: SsChr5, SsChr6 and SsChr7 from *S. spontaneum* [14] and six R570 hom(oe)ology groups HG5-HG10 [13].

Assembling the genome of a polyploid interspecific hybrid is of especially high value for breeders. The assembly, gene prediction, and annotation provided can bridge long standing gaps of knowledge allowing them a more efficient use of genomic tools. Sugarcane's large autopolyploid genome, predominant clonal propagation, and need for extensive phenotyping to determine breeding values, have contributed to the relatively slow (~1% per year at most) rate of progress in improvement of sugarcane [58] and perhaps other autopolyploids. The demonstration that most of its many homo(eo)logs are expressed, often with tissue-specificity, and that transcription factor binding sites and TE insertions differ among homo(eo)logs, suggests complex constraints that may necessitate unusual richness of information to make effective decisions about selecting some homo(eo)logous alleles at the expense of others in autopolyploid breeding populations. These principles may apply widely to many plants with large polyploid genomes that include many of those most efficient at converting solar radiation to biomass.

The present work discloses a large collection of gene-space homo(eo)logs diversity, taking advantage of novel sequencing technologies, adding over 3Gb of sequence not previously reported, in addition to genome annotation, data mined homo(eo)logs, and explored regulatory regions. The presented gene-space of the sugarcane genome is a fundamental step towards a high-quality chromosome resolved assembly from a current commercial hybrid. The genome sequence released for this interspecific polyploid supports its recent allotetraploid nature, reveals differences in promoter regions associated to a diverse gene expression pattern and transposable elements contributing to fine tuning of the sugarcane genome.

METHODS

Plant material

Leaves from SP80-3280 were collected and frozen in liquid nitrogen. Genomic DNA was extracted using DNeasy Plant Mini Kit (Qiagen) following the standard protocol. DNA integrity was analyzed using the Agilent High Sensitivity DNA Analysis Kit (Agilent Technologies) and Agilent 2100 Bioanalyzer Instrument.

Quantification was done using Quant-it™ PicoGreen® dsDNA Assay Kit (ThermoFisher Scientific) and SpectraMax M2 microplate reader (Molecular Devices).

Sequencing Illumina Long-reads and Assembly

We used Illumina Synthetic Long-read sequencing technology, which provides very accurate long reads with a mean read length of roughly 5 kb, thus being able to represent polymorphisms across all copies of chromosomes. Genomic DNA was sheared into 5-10 kb fragments and diluted in 384-well plates. DNA fragments were ligated with PCR primers and specific sequences, which identify the 5' and 3' ends. The fragments from each well were amplified, fragmented and barcoded with unique indices to create 26 TruSeq Synthetic Long-Read DNA libraries. The short fragments created in the second step of fragmentation were pooled and sequenced on the HiSeq instrument at the Illumina Service Genome Network. The reads from each of the 384 wells were pre-processed to correct sequencing and PCR errors. Contigs were produced from the paired-end information and further scaffolded together to resolve repeats and fill in gaps. More details on the informatics pipeline for short read scaffolding into long reads are available in the Fast Track Services Long Reads Pipeline User Guide [59].

To assemble sequences we used a two step approach: *i)* the Celera Assembler [60] (CA) was used for overlap computation and layout building; *ii)*the *tig-sense* module of the HBAR-DTK (Hierarchical-Based AssembleR Development ToolKit) from Pacific Biosciences [61] was used to construct consensus sequences. This was motivated by the fact that the CA, which uses the overlap-layout-consensus method, is more robust than *de Bruijn* graph approaches. However, some adjustments needed to be made. CA, designed for Sanger reads, only accepts quality scores between 0 and 40. Since synthetic long reads are very accurate and some of the base qualities exceeded this upper bound, we transformed the quality scores of our long-read data (FASTQ file) to allow them to be appropriately parsed. The consensus module was also adapted for the analysis of big complex genomes. The substantial number of contigs generated initially (roughly 450,000, half of them singletons) resulted in several files in a folder that hindered I/O operations. So, we *i)* modified *tig-sense* to automatically create subdirectories that contained not more than a thousand contig FASTA files, reducing delays for file lookup; *ii)* divided contig processing into non-singletons and singletons, prioritizing non-singleton contigs;

and *iii)* created a work history so that the program could be resumed after a halt. Overall, these modifications allowed us to reduce the running time of the consensus pipeline by one or two orders of magnitude.

Sequencing BAC clones and assembly

A total of 780 independent BACs were sequenced using Roche454 sequencing technology. Each BAC clone was tagged with a unique barcode and sets of 12 BACs were pooled in one gasket. We assembled BACs individually as described [62] and obtained a total of 49.6 Mbp of assembled sequence, with a mean length of 107 Kbp. The BAC data includes 317 R570 BACs [62], 116 additional R570 BACs and 347 from SP80-3280.

Assembly Validation

Comparison with Sugarcane BACs

Assembled contigs were aligned against a set of 780 BACs with BWA mem, using default parameters. Alignment data was processed for coverage with the aid of Samtools (v1.1) and Bedtools (v2.25) and selected matches were at least 10 kbp long and covered 90% or more of the contig. Additionally, the unassembled synthetic long reads were aligned to the same set of BACs, to check for discrepancies among contigs and long reads, which could be indicative of regions that were not assembled.

Comparison with Sorghum CDS

The set of 39,207 annotated sorghum coding sequences (CDS), release version v2.1, were downloaded from Phytozome [63]. These were aligned against the assembled contigs with BLASTn (v2.2.30+) using default parameters. For each sorghum CDS, we identified the longest fraction of the coding sequence contained within a single unitig. Only hits with at least 80% identity at the nucleotide level were considered for computing coverage.

Comparison with CEGMA

A total of 248 Ultra-conservative core eukaryotic genes classified by Korf Lab [22] were assessed in our sugarcane assembly with '-g' and other default options of CEGMA v2.5. To access the presence of putative homo(eo)logs for CEGMA regions identified on the assembly, the sequences were retrieved according to the

coordinates provided on CEGMA output. Sequences were aligned back to the genome using BLASTn with default parameters. Matches with identity and query coverage greater than 90% were considered for calculation of alignment frequency.

Comparison with BUSCO

Assembly completeness was assessed by searching for the 1,440 core genes from the Plantae lineage of Benchmarking Universal Single-Copy Orthologs (BUSCO) [23]. BUSCO performs gene prediction and orthogonality assessment using Augustus [64] and HMMER3 [65]. Since these steps demand huge resources, we partitioned sugarcane contigs (4.3Gbp) into six groups with similar length and processed BUSCO in parallel. After we merged results, we applied orthogonality assessment algorithm once again as thresholds that BUSCO exploits to discern actual single copy orthologs from paralogs.

Comparison of the mitochondrial and chloroplast genomes

To reconstruct the SP80-3280 mitochondrial and chloroplast genomes, we have used as reference the complete genomes of *Saccharum* hybrid chloroplast (NC_005878.2) [24] and the *Saccharum officinarum* mitochondrial chromosome 1 (LC107874.1) and chromosome 2 (LC107875.1) [25], downloaded from NCBI. The SP80-3280 genome contigs were aligned using BLASTn against their respective references and the best hits were selected based on cutoff E-value ≤ 1x10-15, with contig coverage ≥ 90% and identity ≥ 70%. The BLASTn alignment results identified 2,482 and 909 contigs for the two mitochondrial chromosomes, respectively; and 51,768 contigs for the chloroplast genome. To reconstruct the consensus sequences and do the genome annotation we have used the CLC Genomics Workbench tools [66]. The contigs used for genomes reconstruction presented mean size of 4Kb, with coverage depth higher than 20x.

Using the CLC Tools and the Genome Finishing Module, the selected contigs were aligned to their respective references and consensus sequences extracted, filling the gaps with N's. The reconstructed consensus sequence aligned against the chloroplast genome presented 99.994% and 99.998% of coverage and identity respectively, and there were identified only 6 mismatches and 2 gaps, most of them located in intergenic regions and in one of the rRNA23S copies with protein frame preservation.

The alignment against mitochondrial chromosomes 1 and 2 presented 99.850% and 99,927% of coverage and 99,900% and 99,936% of identity, respectively. The consensus sequences were annotated using their respective NCBI references with the CLC tool "Annotate from Reference", where all genes, tRNAs, rRNAs and miscellaneous features were totally transferred. For the mitochondrial chromosome 1, 237 mismatches and 63 gaps were identified, most of them present in intergenic regions and only 2 mismatches in 2 rRNA genes, with proteins frame preservation. And for chromosome 2, we identified a region composed by 19 N's inside a repetitive AT's region. In addition, the reconstructed chromosome has 57 mismatches and 16 gaps, all of them present in intergenic regions.

Comparison with Sugarcane ESTs

A set of 134,840 ESTs from leaves, internodes and roots samples exclusively from SP80-3280 [20] were aligned to the contigs sequences using SPALN [67] applying mapping and alignment algorithm (-Q 5) and admitting all possible matches for each sequence (-M 1000). Coordinates of aligned ESTs were compared to gene annotation using Bedtools intersect utility [68]. Alignments might be explored through a GBrowse envirioment available at http://sucest-fun.org/cgi-bin/cane_regnet/gbrowse2/gbrowse/microsoft_genome_moleculo_scga7/ or http://sucest-fun.org/wsapp/ > Cane Genome > Sugarcane Genome Microsoft-Moleculo).

Genome Annotation

Gene prediction

Contigs were annotated using a pipeline developed in house, previously used for BAC annotation. Transposable element (TE) discovery and masking was done using LTR harvest, LTR digest, CrossMatch against *Utricularia gibba* TE DB and RepeatMasking [69] of Viridiplantae [70] and previously known sugarcane TEs [47].

Genes were discovered and annotated using masked contig sequences. *De novo* predictions were done with Augustus [64], Glimmer HMM [71], GeneMark HMM [72], SNAP and PASA [73] with rice models and sugarcane EST and RNAseq data. Alignments were also generated against reference protein DBs (sorghum, known sugarcane and Phytozome) using Exonerate [74] and BLAST [75] (v2.2.30+). Both *de novo* and

alignment evidence were used for consensus annotation with EVidenceModeler [76] with greater weight given to experimental and alignment information. Functional assignment was derived from protein DB best hits and InterProScan 5 [77] results.

GeneOntology annotation

For functional annotation of predicted proteins from SP80-3280, all sequences were aligned to UniRef50 clusters, a dataset of representative sequences clustering high similarity proteins from UniProtKB [29], using BLASTp (v2.2.30+, *-evalue 1x10-5*). Sequences that fail to align in this first approach were also searched against the RefSeq non-redundant protein database. Gene Ontology mapping and annotation of sequences with positive BLAST results was performed using Blast2Go framework [78].

Reference-guided RNAseq Assembly

We used Trinity version 2.0.6 for reassembly of the Sugarcane ORFeome [28] using the genome as a reference, with a minimum contig length of 250 bp (genome_guided_max_intron 3,000, genome_guided_min_coverage 5, genome_guided_min_reads_per_partition 10) to identify transcript models. SP80-3280 RNA-seq reads from 3 tissues (leaves and immature and intermediate internodes) were used for alignment against the reference genome and partitioned into read clusters, which were then individually assembled using Trinity genome-guided methods. Trinity and genome-guided methods used a fixed k-mer size of 25nt. In this new assembly, 269,050 genes and 275,807 transcripts were recovered. The quantity of transcripts recovered by the reference guided-assembly was higher, and thus closer to the number of predicted genes (374,774), than the *de novo* assembly. Transcript expression level was estimated by FPKM (fragments per kilobase of exon model per million reads mapped).

Identification of Putative Homo(eo)logs and Count Estimation

We downloaded the *Sorghum bicolor* genome assembly v2.1 from Phytozome and took 2,051 single copy genes according to Han *et al*. [79], which were also present as single copies in the genomes of *Oryza sativa* and *Brachypodium distachyon*. We aligned the coding sequences of these sorghum genes to the coding sequences of predicted sugarcane genes from the SP80-3280 assembly, using the BLASTn (v2.2.30+, *-evalue*

1x10-6). We filtered alignments with at least 80% nucleotide identity, based on Wang *et al*. [50], covering at least 70% of both the sugarcane and sorghum sequences. Sugarcane gene models aligned to the same single copy sorghum gene were denoted as putative homo(eo)logs. Finally, we counted the number of copies for each gene.

We clustered all putative homo(eo)logs based on each single copy sorghum gene to get estimates of sequence differentiation. We aligned the coding sequences for each pairwise combination in each gene cluster, using BLAT v35 [80] (*–minIdentity=0 –minScore=60*). One of the clusters had 21 putative homo(eo)logs, which is higher than the number of chromosome copies expected for sugarcane and was discarded from the analysis. Next, we parsed the alignments to obtain estimates of copy differentiation considering both SNPs and INDELs. We gathered distance estimates from all pairs, from all clusters, to obtain dissimilarity distributions.

Putative Homo(eo)logs characterization

Upstream region analysis

We also assessed the dissimilarity levels of regions upstream (potential promoter regions) of the predicted sugarcane putative homo(eo)logs. We initially collected three different sequence ranges (100 bp, 500 bp and 1,000 bp) upstream of the predicted gene start site. Next, we aligned these upstream sequences for each pairwise combination in each cluster, again using BLAT v35 [80] (*–minIdentity=0 –minScore=30*). Finally, for each distance range, we parsed the alignments and computed the dissimilarity level considering both mismatches and gaps. To avoid partial alignments of the upstream sequences, only alignments up to 20% shorter or longer than the expected sequence length were considered.

Insertions and Deletions between gene copy Coding Sequences

To investigate the occurrence of frameshift mutations between putative homo(eo)logs, we built multiple alignments of its coding sequences for each cluster, with MUSCLE v3.8.31 [82], using default parameters. We then computed the length distribution of insertions and deletions in the coding sequences, to differentiate between frame-preserving and frameshift indels. We parsed the CDS alignment for each pairwise combination of putative homo(eo)logs and counted the number of occurrences of gaps of a given length. We then pooled counts from all copy combinations to get a joint estimated distribution.

573

574 ***Tissue-Specific Homo(eo)logs Expression Analysis***

575 We used RNA-Seq data [28] from leaves (*L*), immature (*I1*) and intermediate (*I5*) internodes of SP80-3280 to
576 find the expression of putative tissue-specific putative homo(eo)logs. These reads were initially aligned to the
577 sugarcane genome assembly using TopHat2 [83] version 2.0.9 (*library-type fr-firststrand*). We allowed reads
578 to be aligned to up to 20 contigs of the genome assembly to identify alignments to different homo(eo)logs (--
579 *max-multihits 20*) and supplied TopHat2 with the putative homo(eo)logs' annotation as a GTF file (--*GTF*
580 *CDSMapping-homo(eo)logs.gtf*), in order to direct TopHat2 to align the reads to this transcriptome first.
581 Besides the *TopHat2* alignment, we used the RSEM tool rsem-calculate-expression (version 1.2.31) to quantify
582 the expression of predicted genes (bowtie2, fragment-length-mean, fragment-length-sd and calc-ci
583 parameters). An in house Perl script was used to estimate the mean length and standard deviation for each
584 RNA-seq library. The main output of *Tophat2 BAM* formatted file [84] *accepted_hits.bam* was used with
585 *RSEM* to estimate the transcriptome expression profile. We developed in-house Perl and R language (version
586 3.3.2) scripts to find the number of putative expressed homo(eo)logs for each single copy Sorghum gene, using
587 the information from *genome annotation* file (GFF format), showing the gene structure, the transcriptome
588 annotation and respective TPM (Transcript Per Million) abundance. The previous information allowed the
589 creation of the homo(eo)logs GFF file. We also applied TopHat2 to find the number of putative homo(eo)logs
590 expressed only in *antisense* orientation, using the same protocol described above, and the *antisense* reads of
591 RNA-Seq previously identified by Nishiyama *et al.* [28].

592

593 **ScSuSy and ScPAL gene family analysis**

594 We used the sugarcane and sorghum SuSy protein sequences reported by Zhang et al. [34] as query for a
595 BLASTx (v2.2.30+) search in the predicted proteins from SP80-3280, *S. spontaneum*[46] and R570 genome
596 assemblies [13]. Putative SuSy genes were then filtered by query coverage >=80% of at least one of the five
597 ScSuSy from Zhang et al. [34] and by PFAM [85] domain search, considering only those containing both the
598 conserved sucrose synthase and glucosyl-transferase 1 domains.
599 Based on BLAST and keyword search ('Phenylalanine ammonia-lyase', 'PAL' and 'EC:4.3.1.24') in two
600 databases (Plant GDB, http://www.plantgdb.org/ and Phytozome [63]) we found 8 different PAL genes in the

sorghum genome, the same number previously reported [86]. For sugarcane, PAL genes were retrieved from an EST Cell Wall catalogue [43], which was used as query together with sorghum PAL genes for a BLASTx (v.2.2.30+) search to identify PAL genes in the predicted proteins from *S. spontaneum* [51] and R570 genome assemblies [13]. Putative PAL genes were then filtered by query coverage ≥ 80% of the sorghum PAL genes and by PFAM [85] domain search, considering only those containing the Aromatic amino acid lyase domain. Also, sequences not containing the PAL conserved amino acid motif Ala-Ser-Gly [87,88] and an essential Tyr110 [89] were excluded.

For both SuSy and PAL, nucleotide sequences (CDS) were aligned with clustalw [90] software in MEGA 7.0 [91] and maximum likelihood trees were constructed with default parameters except for 1,000 bootstraps and Gaps/missing data treatment "*use all sites*". Expression heatmap was constructed using log2 transcript per million (TPM) from previous RNA-seq data [28].

Cell wall-related genes

For the identification of cell wall-related genes in the sugarcane genome we used the Sugarcane SAS Cell Wall catalogue [43] as a reference. The search was carried out using tBLASTn (v2.2.30+, *-evalue 1x10-6*). These were manually re-annotated to produce a sugarcane cell wall catalogue with 3,054 sequences, classified in 10 cell wall categories.

Transcription Factor analysis

For the identification and classification of sugarcane predicted proteins into transcription factor (TF) families, we used the classification rules and tools described in GRASSIUS [46]. The search was carried out using HMMER v3.1b1 [92] and all significant HMM hits with *e*-value smaller than 1x10-3 were kept.

Promoter region analysis

Transcription Start Site (TSS) and promoter region classification

We evaluated promoter regions of genes associated with cell wall and sugar metabolism, ScPAL (Sugarcane Phenylalanine ammonia-lyase) and ScSuSy (Sugarcane Sucrose Synthase), respectively, as described above. A total of 47 ScPAL and 44 ScSuSy was used. To extract the candidate promoter region, we selected, when

available, up to 1,500 nt upstream from the annotated start position of the gene, consisting of a core promoter (500 nt upstream of the start position) and proximal promoter (1,000 nt upstream of the core promoter). Next, we used TSSPlant [93] to predict the TSS of the genes and the type of promoter (TATA-box, TATA-less). The software was set to report high score, sense only TSSs.

Transcription Factor Binding Site (TFBS) in silico characterization

The annotation of TFBSs in the proximal promoter regions was performed in two steps: *de novo* prediction of TFBS motifs in smaller subsets of sequences and mapping the predicted TFBSs in the remaining promoter sequences. Sequences were partitioned in 10 subsets: five ScPAL groups and five ScSuSy groups. We then applied MEME [94] and MotifSampler [95], with default parameters, to each of these datasets to determine putative TFBS motifs. Both were restricted to search for at most 6 motifs with 10nt or less. MEME candidates were a subset of MotifSampler's. MotifSampler ran for 100 cycles; following the manual we selected, from the 10 top-ranked motifs, the first 5 that occurred at least 10 times in the different cycles. Each of the resulting 35 candidate motifs was searched in the JASPAR public database [96], with partial positive matches for all of them.

To evaluate the significance of the motifs we measured their frequency in promoter regions of each of the original gene families and compared them with the frequency of each of these motifs in the promoter regions of the other SP80-3280 predicted genes. We also mapped the motifs of each ScSuSy and ScPAL gene family respectively in the promoter region of the ScSuSy and ScPAL genes from *S. spontaneum* and R570. Candidate motifs were mapped with MotifLocator [95]. For characterizing background sequences, we trained a first order Markov chain [95] trained on SP80-3280 coding regions that were previously shuffled using the fasta-shuffle-letters tool [94]. The parameters were set to full match of the motif in the target sequence and score 95% above of the background.

Co-expression analysis

A field experiment was conducted at the Agricultural Sciences Center of the Federal University of São Carlos in Araras (22°21’25’’S and 47°23’3’’W) in the state of Sao Paulo, Brazil. Trial plots of SP-3280 consisted of four rows of 10 m long and spaced 1.35m apart. The field experiment was initiated in October 2012 and

extended up until November 2013, representing the conditions under which "one-year" sugarcane crops are cultivated. Aiming to carry out observations throughout growth and development, tissue samples of the +1 leaves (L1) and upper (I1), immature (I5) and mature (I9) internodes were collected from two plots (two technical replicates) after 4, 8, 11 and 13 months of planting.

RNA was extracted for four biological replicates, two from each plot, using the TriZol method, treated with DNase I and purified. A pool of samples from leaves and a pool of internodes was used as a 'reference sample' for hybridization experiments on a customized 4×44 K oligoarray (Agilent Technologies) for sugarcane (CaneRegNet), conducted following the recommendations proposed by Lembke et al. [97]. The oligoarrays were read using the GenePix 4000B scanner device (Molecular Devices) and the fluorescence data was processed by Feature Extraction software 9.5.3 (Agilent Technologies).

Log2 transformed expression data was used for discovery and the analysis of co-expression modules, on CEMiTool R package [97]. The adjacency matrix was calculated by estimating the Spearman's correlation coefficient between all pair of genes and raised to a soft thresholding power (β) of 14. TopGO R package [98] was used for gene ontology enrichment analysis for each module and node and edge files were generated for use with the Cytoscape network visualization program [99].

SNP variants (SNVs) analysis compared to genic regions in *Sorghum bicolor*

The 450,609 sugarcane contigs (183,322 singletons and 267,287 unitigs) were aligned to the sorghum genome sequence [52] using the BWA MEM v0.7.10 [100] and contigs with mapping quality larger than 20 were used for variant calling. SNVs were called using samtools v1.1 and bcftools v1.1 [84]. Using in-house Python scripts, extracted SNVs were screened when sugarcane contigs were located on the genic regions of the sorghum genome and two or more sugarcane contigs were aligned to the same sorghum gene. Then, the number of SNVs in each gene was counted according to four-base changes.

SNVs that are homozygous in sugarcane were extracted for further analysis. SNVs mapping to coding regions, splicing sites, stop codons and transcription initiation sites were classified as potential large-effect SNVs.

Functional Enrichment Test

Arabidopsis GO-slim gene annotation was used for functional enrichment analysis. GO-slim terms were assigned to sugarcane genes based on sequence similarity inferred from best BLASTp (v2.2.30+) hit. We used a binomial distribution based on the proportion of a GO-slim term among all annotated genes in the sorghum genome as the null distribution. The binomial test was used to assess functional enrichment, with a significance threshold of $p > 0.05$.

Conserved Synteny Blocks

DNA sequences for all CDSs from *S. spontaneum* [51], R570 [13], *S. bicolor* [101] and SP80-3280 were aligned using the BLASTn program. Results from BLAST searches, with e-value <= 10^{-5}, were parsed using an in-house Python script to filter alignments covering at least 70% of the length of both the query and hit sequences. A second filter, requiring at least 80% identity was also applied and the resulting pairs of queries and hit sequences were classified into putative orthologous groups using the union-find algorithm. We selected putative orthologous groups present in all three organisms but with only one *Sorghum* gene to be used as markers to detect blocks of conserved gene order (syntenic bocks) in comparisons of SP80-3280 and *S. spontaneum* against the genome of *S. bicolor*, thus avoiding the complications of a direct comparison of the two polyploid genomes (**Additional file 1: Fig. S9**). Another Python script was used to detect the syntenic blocks in both *Saccharum* genomes and to count the number of syntenic blocks in each contig. In order to evaluate the effect of genome fragmentation on our estimates of gene conservation, a Monte Carlo simulation of chromosome fragmentation was performed on the R570 and *S. spontaneum* genomes. We sampled 10,000 random regions of the R570 and *S. spontaneum* genomes, with fragment lengths constrained to follow the distribution of contig lengths observed for SP80-3280. We performed 1,000 rounds of these simulated fragmentations, every time allowing genomic fragments (and the genes within them) to be chosen randomly throughout the genome, with no bias to marker genes. We accessed the degree of conservation through the fraction of contigs with two or more marker genes that were found in the same order in the *Saccharum* genome fragments and in the *S. bicolor* genome.

Chromosome Synteny Multiple Correspondence Analysis with Clustering

We performed a multiple correspondence analysis (MCA) with clustering of the best local alignment hit of masked contigs. Input data were the 450,609 contigs of the sugarcane synthetic long read assembly and the

masked genomic sequences of *S. spontaneum* [51] and R570 [13]. We used the masked sugarcane contig sequence produced by the annotation pipeline, excluding 69,879 sequences that were fully masked.

The contigs were aligned to the grass genomes using BWA-SW v0.7.12-r1044 [100]. We used an in-house Perl 5 script to retrieve the highest scoring hit for each contig and generate a table for input into R v3.2.1 [81]. This table contained the chromosome hit, if any, for each contig against each reference genome.

We then used the FactoMineR R package v1.31.3 [102], along with the missMDA missing data handling auxiliary package v1.8.2 [103]. We performed MCA with these data, *i.e.*, chromosome hit number information for each contig was treated as a set of categorical variables and represented in the two principal component dimensions. This was followed by hierarchical clustering in these two dimensions, as well as figure rendering, using the Hierarchical Clustering on Principal Components (HCPC) function of FactoMineR.

In order to identify the correspondence between *S. spontaneum* and R570 chromosomes and SP80-3280 clusters, protein sequence alignment between the cultivar variety and the ancestor and R570 was performed with BLASTp considering an e-value threshold of 1x10-5. The best hit with a minimum query coverage of 90% was selected for visual representation of the alignment results with Circos plot.

ADDITIONAL FILES

Additional file 1.doc contains Supplemental Figures S1 to S11

Additional file 2.xls contains Supplemental Tables S1 to S13

DECLARATIONS

List of abbreviations

CEGMA: Core Eukaryotic Genes Mapping Approach

BUSCO: Benchmarking Universal Single-Copy Orthologs

ESTs: expressed sequence tags

CDS: coding sequences

SuSy: Sucrose Synthase

ScSuSy: Sugarcane Sucrose Synthase

PAL: Phenylalanine ammonia-lyase

ScPAL: Sugarcane Phenylalanine ammonia-lyase

CCR: Cinnamoyl-CoA reductase

COMT: Caffeic acid 3-O-methyltransferase

4CL: 4-coumarate-CoA ligase

TFBSs: Transcription Factor Binding Sites

TE: transposable elements

MCA: Multiple correspondence analysis

I2C-2: R gene locus

dog1: (delay of germination 1

ABRE: ABA-responsive elements

ABA: abscisic acid

Consent for publication: Not applicable

Availability of data and material

Genomic data is publicly available at NCBI under GenBank Bioproject PRJNA431722. Contig sequence, gene annotation, alignment with RNA-seq reads and SAS are also available in a genome browser framework at http://sucest-fun.org/cgi-bin/cane_regnet/gbrowse2/gbrowse/microsoft_genome_moleculo_scga7/ or http://sucest-fun.org/wsapp/ (>Cane Genome >Sugarcane Genome Microsoft-Moleculo). The microarray data have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE124990. All data and scripts are also available at GigaDB and in a Github repository (https://github.com/sp80-3280-genome).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by State of São Paulo Foundation and Microsoft Research (FAPESP grant n° 2012/51062-3) and State of São Paulo Foundation (FAPESP grants n° 2014/50921-8, 2008/52146-0 and 2008/52074-0) under the BIOEN Program. Additional funding included awards from the National Science Foundation (DBI-1350041), and from the National Institutes of Health (R01-HG006677). Bioinformatic tools were run locally on the servers HELIX -IQ / Lab. Signal Transduction - and on the eScience Network - IME / FAPESP grant n° 2011 / 50761-2, CNPq, CAPES, NAP eScience - PRP – USP.

GMS is a recipient of a CNPq Productivity Fellowship 304360/2014-7; MAVS is a recipient of a CNPq Productivity Fellowship (308197/2010-0); GRAM was supported by the FAPESP grant 2015/22993-7; JW was supported by the FAPESP Fellowships 2013/18322-4 and 2015/15346-5 and CNPq Fellowship 159094/2014-3; ALD is a recipient of a FAPESP Fellowship 2017/02270-6; MMO was a recipient of a CAPES Fellowship DS-1454337; SSF was supported by the FAPESP Fellowships 2013/23048-9 and 2016/06917-1; MYN was supported by a FAPESP fellowship 2013/07467-1; FTC is a recipient of a FAPESP Fellowship 2017/02842-0; AMD is a recipient of a CNPq Productivity Fellowship (309566/2015-0); AP is a recipient of funding from the International Consortium for Sugarcane Biotechnology; US National Science Foundation IOS-0115903, and Georgia Agricultural Experiment Station.

Authors' contributions

Project leaders: GMS, MAVS and DH;

Sample collection and DNA extraction: CGL;

Genome sequencing and assembly: HL, MCS, GRAM, RP and BD;

Genome assembly supervision: DH;

Genome annotation: MAVS, GJW, MYNJ and FTC;

Saccharum spontaneum genome assembly: JZ, XZ, QZ and RM;

BWA-SW analysis: GJW;

BAC sequencing and assembly: MAVS, GJW, GTR, HB and SV;

Synteny analysis: AMD, RFS and GGS;

Reference-guided RNAseq Assembly: MYNJ;

Tissue-Specific Allelic Expression Analysis: MYNJ, CGL and PMA;

Phylogeny analysis: SSF and ALD;

SP80-3280 growth and maturation experiment: MSC, GMS, CGL and ALD

Co-expression analysis: ALD

Regulatory region analysis (TE and TFBS): MAVS, MMO, AMD, GMS, CTH and ALD;

SNP variants (SNVs) analysis: CK, HG and AP;

Organization and management of the author's contributions: CGL, ALD, GMS and MAVS;

Data availability (NCBI, Github and Sucest-fun): FTC;

All authors have read and approved the final version of the manuscript.

Acknowledgements

We are indebted to Andreia Prata, Vania Sedano, Nathalia de Setta, Joni Lima, Marcos Buckeridge, Eveline Tavares, Katia Scortecci, Anete Pereira de Souza, Sonia Vautrin and Hélène Bergès for contributions in BAC library construction, BAC selection or sequencing. We are indebted to the Sugarcane Genome Sequencing Initiative for useful discussions.

REFERENCES

1. FAOSTAT. Production/Crops, Food and Agriculture Organization of the United Nations - Statistics Division [Internet]. 2018. Available from: http://www.fao.org/faostat/en/#home

2. Long SP, Karp A, Buckeridge SC, Davis SC, Jaiswal D, Moore PH, et al. Feedstocks for biofuels and bioenergy. Bioenergy Sustain Bridg Gaps [Internet]. Paris Cedex: Scientific Committee on Problems of the Environment (SCOPE); 2015. p. 302–347. Available from: http://bioenfapesp.org/scopebioenergy/images/chapters/bioen-scope_chapter10.pdf

3. Kline KL, Msangi S, Dale VH, Woods J, Souza GM, Osseweijer P, et al. Reconciling food security and bioenergy: priorities for action. GCB Bioenergy. 2017;9:557–76.

4. Goldemberg J. Ethanol for a Sustainable Energy Future. Science. 2007;315:808–10.

5. Jaiswal D, De Souza AP, Larsen S, LeBauer DS, Miguez FE, Sparovek G, et al. Brazilian sugarcane ethanol as an expandable green alternative to crude oil use. Nat Clim Change. 2017;7:788–92.

6. Souza GM, Ballester MVR, de Brito Cruz CH, Chum H, Dale B, Dale VH, et al. The role of bioenergy in a climate-changing world. Environ Dev. 2017;23:57–64.

7. Souza GM, Victoria RL, Joly CA, Verdade LM. Bioenergy & sustainability: bridging the gaps. Paris Cedex: Scientific Committee on Problems of the Environment (SCOPE); 2015.

8. Souza GM, Filho RM. Industrial Biotechnology and Biomass: What Next for Brazil's Future Energy and Chemicals? Ind Biotechnol. 2016;12:24–5.

9. Vilela M de M, Del-Bem L-E, Van Sluys M-A, de Setta N, Kitajima JP, Cruz GMQ, et al. Analysis of three sugarcane homo/homeologous regions suggests independent polyploidization events of *Saccharum officinarum* and *Saccharum spontaneum*. Genome Biol Evol. 2017;evw293.

10. Jannoo N, Grivet L, Seguin M, Paulet F, Domaingue R, Rao PS, et al. Molecular investigation of the genetic base of sugarcane cultivars. Theor Appl Genet. 1999;99:171–84.

11. D'Hont A. Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. Cytogenet Genome Res. 2005;109:27–33.

12. Thirugnanasambandam PP, Hoang NV, Henry RJ. The Challenge of Analyzing the Sugarcane Genome. Front Plant Sci [Internet]. 2018 [cited 2018 Aug 23];9. Available from: http://journal.frontiersin.org/article/10.3389/fpls.2018.00616/full

13. Garsmeur O, Droc G, Antonise R, Grimwood J, Potier B, Aitken K, et al. A mosaic monoploid reference sequence for the highly complex genome of sugarcane. Nat Commun [Internet]. 2018 [cited 2018 Aug 16];9. Available from: http://www.nature.com/articles/s41467-018-05051-5

14. Zhang J, Zhang X, Tang H, Zhang Q, Hua X, Ma X, et al. Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. Nat Genet. 2018;50:1565–73.

15. Waclawovsky AJ, Sato PM, Lembke CG, Moore PH, Souza GM. Sugarcane for bioenergy production: an assessment of yield and regulation of sucrose content. Plant Biotechnol J. 2010;8:263–76.

16. Goldemberg J, Coelho ST, Guardabassi P. The sustainability of ethanol production from sugarcane. Energy Policy. 2008;36:2086–97.

17. Welbaum GE, Meinzer FC. Compartmentation of solutes and water in developing sugarcane stalk tissue. Plant Physiol. 1990;93:1147–53.

18. Bonawitz ND, Chapple C. The genetics of lignin biosynthesis: connecting genotype to phenotype. Annu Rev Genet. 2010/09/03. 2010;44:337–63.

19. Himmel ME, Ding SY, Johnson DK, Adney WS, Nimlos MR, Brady JW, et al. Biomass recalcitrance: engineering plants and enzymes for biofuels production. Science. 2007/02/10. 2007;315:804–7.

20. Vettore AL. Analysis and Functional Annotation of an Expressed Sequence Tag Collection for Tropical Crop Sugarcane. Genome Res. 2003;13:2725–35.

21. Riaño-Pachón DM, Mattiello L. Draft genome sequencing of the sugarcane hybrid SP80-3280. F1000Research. 2017;6:861.

22. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23:1061–7.

23. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.

24. Calsa Júnior T, Carraro DM, Benatti MR, Barbosa AC, Kitajima JP, Carrer H. Structural features and transcript-editing analysis of sugarcane (Saccharum officinarum L.) chloroplast genome. Curr Genet. 2004;46:366–73.

869 25. Shearman JR, Sonthirod C, Naktang C, Pootakham W, Yoocha T, Sangsrakru D, et al. The two chromosomes
870 of the mitochondrial genome of a sugarcane cultivar: assembly and recombination analysis using long PacBio
871 reads. Sci Rep [Internet]. 2016 [cited 2018 Jan 24];6. Available from:
872 http://www.nature.com/articles/srep31533

873 26. Cardoso-Silva CB, Costa EA, Mancini MC, Balsalobre TWA, Canesin LEC, Pinto LR, et al. De Novo Assembly
874 and Transcriptome Analysis of Contrasting Sugarcane Varieties. Gibas C, editor. PLoS ONE. 2014;9:e88462.

875 27. Vicentini R, Bottcher A, Brito M dos S, dos Santos AB, Creste S, Landell MG de A, et al. Large-Scale
876 Transcriptome Analysis of Two Sugarcane Genotypes Contrasting for Lignin Content. Amancio S, editor. PLOS
877 ONE. 2015;10:e0134909.

878 28. Nishiyama MY, Ferreira SS, Tang P-Z, Becker S, Pörtner-Taliana A, Souza GM. Full-Length Enriched cDNA
879 Libraries and ORFeome Analysis of Sugarcane Hybrid and Ancestor Genotypes. PLOS ONE. 2014;9:e107351.

880 29. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, the UniProt Consortium. UniRef clusters: a
881 comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics.
882 2015;31:926–32.

883 30. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the
884 unification of biology. Nat Genet. 2000;25:25–9.

885 31. Veeckman E, Ruttink T, Vandepoele K. Are We There Yet? Reliably Estimating the Completeness of Plant
886 Genome Sequences. Plant Cell. 2016;28:1759–68.

887 32. Nelson JC, Wang S, Wu Y, Li X, Antony G, White FF, et al. Single-nucleotide polymorphism discovery by
888 high-throughput sequencing in sorghum. BMC Genomics [Internet]. 2011 [cited 2018 Jan 26];12. Available
889 from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-352

890 33. Coleman HD, Yan J, Mansfield SD. Sucrose synthase affects carbon partitioning to increase cellulose
891 production and altered cell wall ultrastructure. Proc Natl Acad Sci. 2009;106:13118–23.

892 34. Zhang J, Arro J, Chen Y, Ming R. Haplotype analysis of sucrose synthase gene family in three
893 Saccharumspecies. BMC Genomics. 2013;14:314.

894 35. Rawat R, Schwartz J, Jones MA, Sairanen I, Cheng Y, Andersson CR, et al. REVEILLE1, a Myb-like
895 transcription factor, integrates the circadian clock and auxin pathways. Proc Natl Acad Sci. 2009;106:16883–
896 8.

897 36. Seo PJ, Ryu J, Kang SK, Park C-M. Modulation of sugar metabolism by an INDETERMINATE DOMAIN
898 transcription factor contributes to photoperiodic flowering in Arabidopsis: Sugar and photoperiodic
899 flowering. Plant J. 2011;65:418–29.

900 37. Papini-Terzi FS, Rocha FR, Vêncio RZ, Felix JM, Branco DS, Waclawovsky AJ, et al. Sugarcane genes
901 associated with sucrose content. BMC Genomics. 2009;10:120.

902 38. Persia D, Cai G, Del Casino C, Faleri C, Willemse MT, Cresti M. Sucrose synthase is associated with the cell
903 wall of tobacco pollen tubes. Plant Physiol. 2008;147:1603–18.

904 39. Brill E, van Thournout M, White RG, Llewellyn D, Campbell PM, Engelen S, et al. A Novel Isoform of Sucrose
905 Synthase Is Targeted to the Cell Wall during Secondary Cell Wall Synthesis in Cotton Fiber. Plant Physiol.
906 2011;157:40–54.

907 40. Sewalt Vjh, Ni W, Blount JW, Jung HG, Masoud SA, Howles PA, et al. Reduced Lignin Content and Altered
908 Lignin Composition in Transgenic Tobacco Down-Regulated in Expression of L-Phenylalanine Ammonia-Lyase
909 or Cinnamate 4-Hydroxylase. Plant Physiol. 1997;115:41–50.

910 41. Rohde A. Molecular Phenotyping of the pal1 and pal2 Mutants of Arabidopsis thaliana Reveals Far-
911 Reaching Consequences on Phenylpropanoid, Amino Acid, and Carbohydrate Metabolism. PLANT CELL
912 ONLINE. 2004;16:2749–71.

913 42. Vanholme R, Storme V, Vanholme B, Sundin L, Christensen JH, Goeminne G, et al. A Systems Biology View
914 of Responses to Lignin Biosynthesis Perturbations in Arabidopsis. Plant Cell. 2012;24:3506–29.

915 43. Ferreira SS, Hotta CT, Poelking VG de C, Leite DCC, Buckeridge MS, Loureiro ME, et al. Co-expression
916 network analysis reveals transcription factors associated to cell wall biosynthesis in sugarcane. Plant Mol Biol.
917 2016;91:15–35.

918 44. Cunha CP, Roberto GG, Vicentini R, Lembke CG, Souza GM, Ribeiro RV, et al. Ethylene-induced
919 transcriptional and hormonal responses at the onset of sugarcane ripening. Sci Rep [Internet]. 2017 [cited
920 2018 Aug 16];7. Available from: http://www.nature.com/articles/srep43364

921 45. Xu Z, Zhang D, Hu J, Zhou X, Ye X, Reichel KL, et al. Comparative genome analysis of lignin biosynthesis
922 gene families across the plant kingdom. BMC Bioinformatics. 2009;10:S3.

923 46. Yilmaz A, Nishiyama MY, Fuentes BG, Souza GM, Janies D, Gray J, et al. GRASSIUS: A Platform for
924 Comparative Regulatory Genomics across the Grasses. PLANT Physiol. 2009;149:171–80.

925 47. Domingues DS, Cruz GM, Metcalfe CJ, Nogueira FT, Vicentini R, de S Alves C, et al. Analysis of plant LTR-
926 retrotransposons at the fine-scale family level reveals individual molecular patterns. BMC Genomics.
927 2012;13:137.

928 48. Kim C, Wang X, Lee T-H, Jakob K, Lee G-J, Paterson AH. Comparative Analysis of Miscanthus and
929 Saccharum Reveals a Shared Whole-Genome Duplication but Different Evolutionary Fates. Plant Cell.
930 2014;26:2420–9.

931 49. Vieira MLC, Almeida CB, Oliveira CA, Tacuatiá LO, Munhoz CF, Cauz-Santos LA, et al. Revisiting Meiosis in
932 Sugarcane: Chromosomal Irregularities and the Prevalence of Bivalent Configurations. Front Genet [Internet].
933 2018 [cited 2018 Aug 27];9. Available from:
934 https://www.frontiersin.org/article/10.3389/fgene.2018.00213/full

935 50. Wang J, Roe B, Macmil S, Yu Q, Murray JE, Tang H, et al. Microcollinearity between autopolyploid
936 sugarcane and diploid sorghum genomes. BMC Genomics. 2010;11:261.

937 51. Zhang et al. Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. Accept Nat
938 Genet. 2018;

939 52. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The Sorghum bicolor
940 genome and the diversification of grasses. Nature. 2009;457:551–6.

941 53. D'Hont A, Ison D, Alix K, Roux C, Glaszmann JC. Determination of basic chromosome numbers in the genus
942 *Saccharum* by physical mapping of ribosomal RNA genes. Genome. 1998;41:221–5.

943 54. Liu Z, Adams KL. Expression Partitioning between Genes Duplicated by Polyploidy under Abiotic Stress
944 and during Organ Development. Curr Biol. 2007;17:1669–74.

945 55. Ramírez-González RH, Borrill P, Lang D, Harrington SA, Brinton J, Venturini L, et al. The transcriptional
946 landscape of polyploid wheat. Science. 2018;361:eaar6089.

947 56. Zhang Y, Liu Z, Khan AA, Lin Q, Han Y, Mu P, et al. Expression partitioning of homeologs and tandem
948 duplications contribute to salt tolerance in wheat (Triticum aestivum L.). Sci Rep [Internet]. 2016 [cited 2018
949 Aug 16];6. Available from: http://www.nature.com/articles/srep21476

950 57. Liu Z, Xin M, Qin J, Peng H, Ni Z, Yao Y, et al. Temporal transcriptome profiling reveals expression
951 partitioning of homeologous genes contributing to heat and drought acclimation in wheat (Triticum aestivum
952 L.). BMC Plant Biol [Internet]. 2015 [cited 2018 Aug 16];15. Available from:
953 http://www.biomedcentral.com/1471-2229/15/152

954 58. Dal-Bianco M, Carneiro MS, Hotta CT, Chapola RG, Hoffmann HP, Garcia AAF, et al. Sugarcane
955 improvement: how far can we go? Curr Opin Biotechnol. 2012;23:265–70.

956 59. Illumina. FastTrack Services Long Reads Pipeline User Guide. 2013.

957 60. Myers EW. A Whole-Genome Assembly of Drosophila. Science. 2000;287:2196–204.

958 61. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial
959 genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10:563–9.

960 62. de Setta N, Monteiro-Vitorello CB, Metcalfe CJ, Cruz GMQ, Del Bem LE, Vicentini R, et al. Building the
961 sugarcane genome for biotechnology and identifying evolutionary trends. BMC Genomics. 2014;15:540.

962 63. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform
963 for green plant genomics. Nucleic Acids Res. 2012;40:D1178–86.

964 64. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein
965 multiple sequence alignments. Bioinforma Oxf Engl. 2011;27:757–63.

966 65. Eddy SR. Accelerated Profile HMM Searches. Pearson WR, editor. PLoS Comput Biol. 2011;7:e1002195.

967 66. Knudsen T, Knudsen B. CLC Genomics Benchwork 6 [Internet]. 2013. Available from:
968 http://www.clcbio.com

969 67. Gotoh O. Direct mapping and alignment of protein sequences onto genomic sequence. Bioinformatics.
970 2008;24:2438–44.

971 68. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics.
972 2010;26:841–2.

973 69. Smit A, Hubley R, Green P. RepeatMasker Open-4.0 [Internet]. Available from:
974 http://www.repeatmasker.org

975 70. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of
976 eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7.

977 71. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic
978 gene-finders. Bioinforma Oxf Engl. 2004;20:2878–9.

979 72. Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and
980 viruses. Nucleic Acids Res. 2005;33:W451–4.

981 73. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, et al. Improving the Arabidopsis
982 genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31:5654–66.

983 74. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC
984 Bioinformatics. 2005;6:31.

985 75. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and
986 applications. BMC Bioinformatics. 2009;10:421.

987 76. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure
988 annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol.
989 2008;9:R7.

990 77. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein
991 function classification. Bioinformatics. 2014;30:1236–40.

992 78. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation,
993 visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–6.

994 79. Han F, Peng Y, Xu L, Xiao P. Identification, characterization, and utilization of single copy genes in 29
995 angiosperm genomes. BMC Genomics. 2014;15:504.

996 80. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12:656–64.

997 81. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2014.
998 Available from: http://www.R-project.org

999 82. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids
1000 Res. 2004;32:1792–7.

1001 83. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of
1002 transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.

1003 84. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format
1004 and SAMtools. Bioinforma Oxf Engl. 2009;25:2078–9.

1005 85. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database:
1006 towards a more sustainable future. Nucleic Acids Res. 2016;44:D279–85.

1007 86. Xu Z, Zhang D, Hu J, Zhou X, Ye X, Reichel KL, et al. Comparative genome analysis of lignin biosynthesis
1008 gene families across the plant kingdom. BMC Bioinformatics. 2009;10 Suppl 1:S3.

1009 87. Röther D, Poppe L, Morlock G, Viergutz S, Rétey J. An active site homology model of phenylalanine
1010 ammonia-lyase from P. crispum. Eur J Biochem. 2002;269:3065–75.

1011 88. Calabrese JC, Jordan DB, Boodhoo A, Sariaslani S, Vannelli T. Crystal structure of phenylalanine ammonia
1012 lyase: Multiple helix dipoles implicated in catalysis. Biochemistry. 2004;43:11403–16.

1013 89. Pilbák S, Tomin A, Rétey J, Poppe L. The essential tyrosine-containing loop conformation and the role of
1014 the C-terminal multi-helix region in eukaryotic phenylalanine ammonia-lyases. FEBS J. 2006;273:1004–19.

1015 90. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple
1016 sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.
1017 Nucleic Acids Res. 1994;22:4673–80.

1018 91. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger
1019 Datasets. Mol Biol Evol. 2016;33:1870–4.

92. Zhang Z, Wood WI. A profile hidden Markov model for signal peptides generated by HMMER. Bioinforma Oxf Engl. 2003;19:307–8.

93. Shahmuradov IA, Umarov RKh, Solovyev VV. TSSPlant: a new tool for prediction of plant Pol II promoters. Nucleic Acids Res. 2017;gkw1353.

94. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202–8.

95. Claeys M, Storms V, Sun H, Michoel T, Marchal K. MotifSuite: workflow for probabilistic motif detection and assessment. Bioinformatics. 2012;28:1931–2.

96. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2018;46:D260–6.

97. Russo PST, Ferreira GR, Cardozo LE, Bürger MC, Arias-Carrasco R, Maruyama SR, et al. CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. BMC Bioinformatics [Internet]. 2018 [cited 2018 Aug 16];19. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2053-1

98. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology.

99. Shannon P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003;13:2498–504.

100. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinforma Oxf Engl. 2010;26:589–95.

101. McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, et al. The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. Plant J Cell Mol Biol. 2018;93:338–54.

102. Lê S, Josse J, Husson F. FactoMineR : An R Package for Multivariate Analysis. J Stat Softw [Internet]. 2008 [cited 2017 Nov 30];25. Available from: http://www.jstatsoft.org/v25/i01/

103. Josse J, Husson F. missMDA : A Package for Handling Missing Values in Multivariate Data Analysis. J Stat Softw [Internet]. 2016 [cited 2017 Nov 30];70. Available from: http://www.jstatsoft.org/v70/i01/

Table 1 – Genome sequencing: Technology and assembly details and gene prediction features.

	Description	Genomic DNA	BAC clones
Sequencing and assembly data	Sequencing Data	26 Illumina synthetic long-read libraries	Single end Roche 454 of BAC library clones
	Total Sequence	19 Gb	6.6 Gb
	Genome coverage	1.9 x	0.66 x
	Read length Min/Max/Mean	1,500 bp / 22,904 bp / 4,930 bp	8 bp / 2611 bp / 368.5 bp
	Assembler Software	Celera Assembler (Overlap Graph)	PHRAP/CONSED
	Total reads used in assembly	3,857,849	17,894,306
	Total assembly size	4.26 Gb	49.6 Mb
	Number of unitigs/contigs + singletons	450,609	463
	Contigs Length Min/Max/Mean	1,500 bp / 468,011 bp / 9,452 bp	11,723 bp / 235,533 bp / 107,129 bp
	NG50	41,394 bp	109,618 bp
	N50	13,157 bp	N/A
Gene prediction features	# genes	373,869	3,550
	# transcripts	374,774	-
	# exons	1,035,764	13,132
	Average GC content	43.20%	44.99%
	Average # exons per gene	2.8	3.7
	Average exon size [bp]	291	271.8
	Median exon size [bp]	171	154
	Average intron size [bp]	352.6	539.2
	Median intron size [bp]	132	139
	Average gene size [bp] with UTR	1,437.80	2,429.20
	Median gene size [bp] with UTR	806	1,260.50
	Average gene size [bp] without UTR	1,318.80	2,351.30
	Median gene size [bp] without UTR	771	1,199.50
	Average gene density (kb per gene)	11.4	14

Figure captions

Fig. 1 – Gene copy number estimation. (**A**) Distribution of copy counts for putative single copy genes. A total of 1,592 single copy genes from sorghum, rice and *Brachypodium* matched sugarcane predicted genes. More than 99.9% of the aligned single copy genes are present between one and 15 times in the sugarcane gene models. (**B**) Copy differentiation between sugarcane coding sequences (CDS) and upstream regions, based on pairwise sequence alignment of gene clusters. Genetic dissimilarity increases with increasing distance from the translation start site. (**C**) Indel length distribution in sugarcane putative homo(eo)logs. Frame preserving indels are more common than frameshifts for this set of genes.

Fig. 2 – Homo (eo)log expression: Frequency of sugarcane genes plotted against the total number of homo(eo)logs per gene and the number of expressed homo(eo)logs per gene. Genes with cDNAs aligned with FPKM > 1 were considered expressed. Plots show sense (**A**) and antisense (**B**) transcripts. Reads from Ion PGM Sequencing were used, as strand orientation is maintained [28].

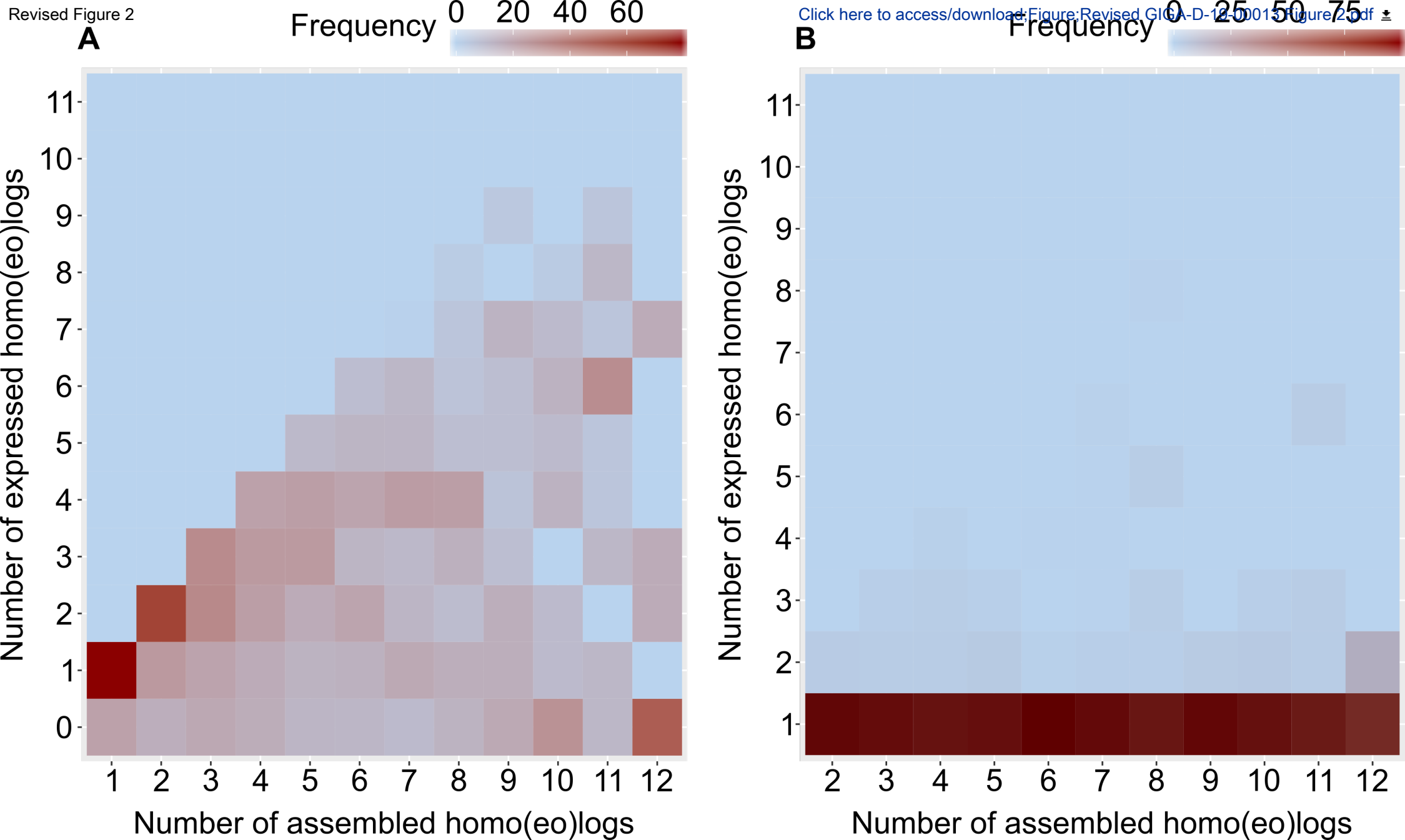
Fig. 3 – Phylogeny, putative regulatory regions and expression of sucrose synthase (SuSy) and phenylalanine-ammonia lyase (PAL) gene family. Phylogenetic analysis of (**A**) SuSy and (**B**) PAL genes from SP80-3280, R570, *S. spontaneum*, and sorghum. SuSy sequences from *Saccharum* ssp [34] were also included. For both SuSy and PAL, nucleotide sequences (CDS) were aligned with CLUSTALW [90] software in MEGA 7.0 [91] and maximum likelihood trees were constructed with default parameters except for 1,000 bootstraps (only bootstraps values over 20 are shown). Core promoter analysis (gray columns in C and D) using TSSPlant [93] suggests ScSuSy2 (**C**) and most ScPAL (**D**) as TATA-less (absence of black squares). Transcription factor binding sites (TFBS) prediction (colored symbols in **C** and **D**) using MEME [94] and MotifSampler [95] suggest specific motif for each group (ScSuSy1, ScSuSy2, ScSuSy5 and PAL I, PAL III, PAL Va and PAL Vb). The three SP80-3280 PAL genes marked (* in **D**) are present in the same contig. Transposable elements (TEs) were identified within 10 kb upstream from the gene (**C** and **D**). Heatmap analysis of RNA-Seq data [28] (expression profile in **C** and **D**) shows more pronounced expression in SP80-3280 internodes (I1 and I5) of ScSuSy1, ScSuSy2, ScSuSy5 and PAL from group V. RNA-Seq of leaf tissues (L) indicates more pronounced expression of ScPAL from groups II and III. ScSuSy3 presents high numbers of TFBS and TE and low expression in all samples.

Fig. 4 – SNP variants. Alignment of sugarcane contigs to the genic regions of sorghum chromosomes (chromosome 1 is on top and 10 is at the bottom). X and Y axes indicate physical distance on each chromosome (mega base pairs, Mb) and the number of single nucleotide variants compared to the sorghum reference genome, respectively. Each dot indicates sorghum genes matching two or more sugarcane contigs.

Fig. 5 – Pseudoassembly of contigs. Multiple correspondence analysis (MCA) with hierarchical clustering of the SP80-3280 assembly against the *S. spontaneum* tetraploid AP85-441 homo(eo)log-resolved assembly [14] and the R570 [13] monoploid genome. **A**: SP80-3280 contigs best hits against AP85-441 and R579 chromosomes and corresponding size of the preliminary scaffolds; Cluster = hierarchical cluster from the MCA. **B** and **C**: Circos plot of the proportion of proteins from SP80-3280 (classified into one of the 6 clusters or as 'non-clustered') that align to the AP85-441 and R570 putative chromosomes, respectively.

Revised Figure 2

Click here to access/download;Figure;Revised GIGA-D-18-00015 Figure 2.pdf



Click here to access/download
Supplementary Material
Revised GIGA-D-19-00013 Additional file 1.docx

Click here to access/download
Supplementary Material
Revised GIGA-D-19-00013 Additional file 2.xls

Response letter

Click here to access/download;Personal Cover;Response letter GIGA-D-19-00013.docx

May 23rd, 2019

GIGA-D-19-00013

Assembly of the 373K gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop

Glaucia Mendes Souza, Ph.D; Marie-Anne Van Sluys, Ph.D; Carolina Gimiliani Lembke, Ph.D; Hayan Lee, Ph.D; Gabriel Rodrigues Alves Margarido, Ph.D; Carlos Takeshi Hotta, Ph.D; Jonas Weissmann Gaiarsa, Ph.D; Augusto Lima Diniz, Ph.D; Mauro de Medeiros Oliveira, Ph.D; Sávio de Siqueira Ferreira, Ph.D; Milton Yutaka Nishiyama-Jr, Ph.D; Felipe ten Caten, Ph.D; Geovani Tolfo Ragagnin, MSc; Pablo de Morais Andrade, Ph.D; Robson Francisco de Souza, Ph.D; Gianlucca Gonçalves Nicastro, Ph.D; Ravi Pandya, BS.c; Changsoo Kim, Ph.D; Hui Guo, Ph.D; Alan Mitchell Durham, Ph.D; Monalisa Sampaio Carneiro, Ph.D; Jisen Zhang, Ph.D; Qing Zhang, Ph.D; Qing Zhang, Ph.D; Ray Ming, Ph.D; Michael Schatz, Ph.D; Bob Davidson; Andrew Paterson, Ph.D; David Heckerman, Ph.D

Dear Dr. Hans Zauner

Assistant Editor

Gigascience

Firstly, we would like to thank the editor and reviewers for the valuable comments on our submitted manuscript. We have responded to all comments and criticisms and have revised the paper in light of them. A point-by-point response to these concerns is provided below. We reinforce the high value of the presented sugarcane hybrid gene space assembly, not only for the sugarcane community, but also for those interested in unraveling genomics of complex polyploid crops. The revised version of our manuscript (in addition to Fig.2 and Additional files 1 and 2) has been uploaded.

Sincerely,

Glaucia Mendes Souza
Full Professor
Institute of Chemistry
University of São Paulo

Marie-Anne Van Sluys
Full Professor
Biosciences Institute
University of São Paulo

Reviewer #1: Souza et al. report a gene-space assembly of the polyploid sugarcane genome. They tackled the difficulty of assembling a highly redundant gene-space by relying on the Illumina Long Read technology and could report an impressive gene count of over 370,000 gene models, which can be expected for a genome that is expected to contain between 8 and 13 sets of chromosomes.

The gene-space assembly will provide an important resource for establishing more functional genomic studies in sugarcane, eventually leading also to new approaches of sugarcane crop improvement through genomics-based crop improvement, most likely, however, through transgenic approaches including gene editing.

The manuscript needs to be corrected for typos: homo(eo)logs genes and homo(eo)logs alleles is a recurring mistake - it should either read homo(eo)logs or homo(eo)logous genes / alleles.

Response: We apologize for these mistakes and we have corrected the term in the revised version.

Reviewer #1: The paper reports a sequence resource which is used for sequence analysis and interpretation of a few gene families - while this is of interest in respect of the economic use of the crop, the analysis remains speculative and descriptive thus could be presented in a more concise way.

Response: We accept the reviewer's suggestion and have revisited the manuscript and reduced accordingly as stated in the next comment.

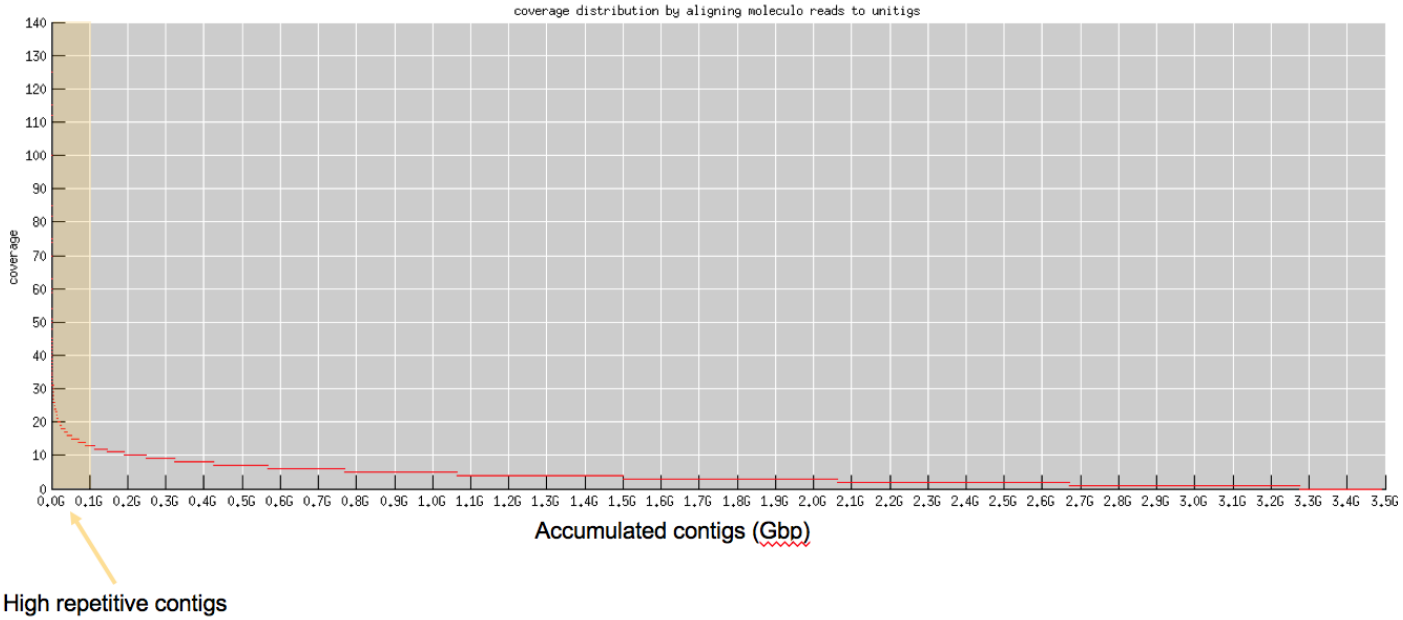
Reviewer #1: motif analysis results and SNV density are reported only or again in the discussion. I highly recommend to shorten the manuscript by merging results and discussion or shorten significantly the discussion and conclusion section, which will increase the readability of the manuscript.

Response: We accept the reviewer's suggestion, have revisited the discussion section and have reallocated the Motif analysis and SNV paragraphs. In addition, we followed the GigaScience instructions for preparing the main manuscript text, which indicatess that the Discussion should be presented in a separate section.

Reviewer #1: I would have appreciated a stronger attempt of using independent evidence for the redundancy of the gene set. Especially in regard of the promoter analysis it would be more intuitive to see an assessment of sequence quality in the 5`and 3`regions. This might be hidden in the supplemental material, which I could not access. It remains elusive to me if the authors have assessed their assembly by a read coverage analysis in order to identify problematic regions in the contigs/unitigs.

Response: We indicate at L426: "… synthetic long reads are very accurate", so much so that we had to modify the quality scores of our long-read data to allow them to be appropriately parsed. In addition, we did assess our assembly by a read coverage analysis in order to identify problematic regions by mapping reads back to contigs. After sorting contigs from highest coverage to lowest, we found that only 0.1 Gbp of contigs had very high coverage which means

a lot of repeats. In addition, there was a reasonable amount of coverage of less than 8-12x (refer to figure below).



Reviewer #1: The authors report the use of custom scripts in their analyses. These must be submitted to a public repository, e.g. Github or others.

Response: We would like to remind that all data and scripts have been submitted to GigaDB following GigaScience submission guidelines. In addition, and accepting the reviewer's suggestion, we have created a Github repository available upon publication (https://github.com/sp80-3280-genome). Other than GigaDB, Github and NCBI, the data will also be available through SUCEST-FUN database: GBrowse environment available at http://sucest-fun.org/cgi-bin/cane_regnet/gbrowse2/gbrowse/microsoft_genome_moleculo_scga7/ or http://sucest-fun.org/wsapp/ (> Cane Genome >Sugarcane Genome Microsoft-Moleculo). These two links are now present in the new manuscript version, as follows:

L758-765:

"**Availability of data and material**

Genomic data is publicly available at NCBI under GenBank Bioproject PRJNA431722. Contig sequence, gene annotation, alignment with RNA-seq reads and SAS are also available in a genome browser framework at http://sucest-fun.org/cgi-bin/cane_regnet/gbrowse2/gbrowse/microsoft_genome_moleculo_scga7/ or http://sucest-fun.org/wsapp/ (>Cane Genome >Sugarcane Genome Microsoft-Moleculo). The microarray data have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE124990. All data and scripts are also available at GigaDB and in a Github repository (https://github.com/sp80-3280-genome)."

Reviewer #1: Table 1 only reports 454 data for BAC clones while the methods mention also PACbio data - please comment.

Response: We apologize for this mistake and have corrected the text as follows:

L438 – "A total of 780 independent BACs were sequenced using Roche454 sequencing technology"

Reviewer #1: To my opinion section "Sugarcane and sorghum polymorphisms support recent allotetraploidy" and Figure 4 require more explanation. Neither the paragraph nor the figure provide any conclusion that would justify the title of this section. I assume the authors interprete the predominant occurance of biallelic SNPs as an indication of allotetraploidy? but is this really unexpected to have a predominance of biallelic SNPs anyway? It is also unclear what is the context of the 4750 SNP between sorghum and sugarcane. Only the SuSy genes or all 30,000 / 300,000 sorghum/sugarcane homologs? Again I couldn`t access the supplements for this section, however, I recommend to present this part in a more tangible manner.

Response: A passage that clarified the issue raised by the reviewer that originally was presented in the discussion section was relocated to the results section. It makes the case how the SNVs biallelic predominance supports allotetraploidy as previously published by two independent groups.

L281-302:

"**Sugarcane and sorghum polymorphisms support recent allotetraploidy and suggest candidate genes for morphological and physiological differences between these taxa**

Despite a common foundation for evolving high sugar content with similar SuSy genes (ScSuSy1-5), sugarcane and closely related sorghum have taken different paths since sharing ancestry. We identified 10,586 natural SNP variations (SNVs) between sorghum and sugarcane 4,140 unique genes, mostly bi-allelic (80.8%), but 6.2% tri-allelic and 0.97% tetra-allelic (**Fig. 4**). The overwhelming predominance of biallelic variations indicates that many sorghum genes are represented by two discernible sugarcane copies, supporting the theory of allotetraploidization shortly after divergence with sorghum ca. 3.8~4.6 MYA [48], creating two sugarcane 'subgenomes'. Recently published results from Vieira et al. [49], demonstrate that sugarcane meiotic chromosomes behave as bivalents, supporting this inference. Autotetraploidization after Saccharum speciation ca. 3.1~3.8 MYA may have further contributed to allelic richness within each sugarcane 'subgenome' -- the preservation of as many as four functionally different alleles at a locus, with cases observed on all except one chromosome (Chr 10 - **Fig. 4**), is consistent with the well-known heterozygosity of sugarcane cultivars and associated susceptibility to inbreeding depression. However, genes for which sugarcane has only one allele are more abundant than 3- or 4-allele loci, perhaps reflecting cases in which a single gene copy is sufficient, or in which occasional exchanges between subgenomes have homogenized multiple homo(eo)logs.

Further, 1,334 SNVs that differentiate sugarcane from sorghum in 585 single copy genes include frameshifts, premature termination, erroneous splicing, loss of stop codons and incorrect translation initiation (**Additional file 1: Fig. S8**, **Additional file 2: Table S11)** in genes significantly enriched in transcription, DNA-dependent cell organization and biogenesis in the nucleus and endoplasmic reticulum (**Additional file 2: Table S12)** comprise a rich slate of candidates for causes of morphological and physiological differences between these taxa."

Reviewer #1: BUSCO analysis asks for the presence of 1400 single copy genes highly conserved among higher organisms. In theory each of the 8-13 sets of chromosomes is expected to carry a full set of BUSCO genes, at least shortly after the polyploidisation event. so, how conclusive is a report of "5.4% of conserved genes could not be identified" in this context?

Response: BUSCO was being used for *de novo* assembly completeness. For the reason of why 5.4% of BUSCO genes are missing, we have two hypotheses. (1) imperfect BUSCO gene profiling: core gene sequence profile is biased to 'several' species and as such may not be representative of 'true' core plant genes. (2) Although we exploited long synthetic reads, it is still a big challenge to assemble one contig per chromosome. So, the gene may be spread to multiple contigs. That is a limitation of the technology in our time. Since we are aware of such circumstances, we further exploited tBLASTn to thoroughly search 78 missing Plantae lineage BUSCO genes and 70 were covered, only 8 (0.5%) of them remaining absent.

Reviewer #2

The manuscript by Souza et al. describes a gene-space assembly of the large, polyploid genome of a sugarcane commercial hybrid cultivar (SP80-3280) of importance to Brazil. The work builds upon an existing draft assembly of SP80-3280, published by Riaño-Pachón & Mattiello, by adding over 3 Gbp of assembled sequence and additional gene/transcript models to those already published.

Through a variety of methods the authors assess the assembly in terms of how well it represents the gene-space of sugarcane. A general assessment of differential homeolog expression is presented followed by a more in-depth look at two gene families (SuSy and PAL) known to be involved in biomass production. The authors then allocate the sugarcane gene-space contigs into clusters with rough correspondence to the chromosomes of Saccharum spontaneum and the modern sugarcane cultivar R570.

My main concern is in the assessment of the completeness of the gene-space assembly; in particular the potential for over-estimation of its completeness. The authors assess this in terms of identifying regions of sequence similarity in the sugarcane assembly using: 1) sorghum transcripts; 2) ultraconserved eukaryotic genes (CEGMA); 3) single-copy orthologous genes (BUSCO); and 4) assembled mitochondrial and plastid genomes. This sugarcane hybrid is expected to contain up to 15 homeologs of any given gene, and as such if the gene-space assembly is to be deemed "complete" it should contain all homeologs present in the genome. If however, the assembly contains only 1 of each homeolog, the approach taken will still report a high percentage of completeness since each query sequence will likely find a good match in the assembly against the one homeolog. This problem is exacerbated if the sequence alignment parameters and filtering are too relaxed, meaning that query sequences are more likely to find hits which is not orthologous. The authors use a set of ~2,000 genes, which are single-copy in other grass species, to identify the sugarcane homeologs in their assembly and report the number of homeologs identified per single copy gene. Fig. 1 shows the authors identified about <=5 homeologs for half the single-copy genes, much fewer than I had expected. The source of this discrepancy is not discussed. While homeologs may not be present in the genome of this hybrid for a variety of reasons, they may also be present in the genome but absent from the assembly due to lack of sequence coverage or have been collapsed by the assembly algorithm. Therefore, such an approach will likely give rise to an over-estimation of the completeness of an assembly.

Response: We appreciate the reviewer's comments and concerns and provide below a point by point response included in the Specific Points that address all raised issues.

Reviewer #2: Specific points to address:

• L113: The assembly is not accessible via GenBank accession QPEU01000000

Response: We appreciate the reviewer's comment and declare that all data was deposited at GigaDB at the time of submission. Access will be publicly available straight after publication. We also have created a Github repository available at https://github.com/sp80-3280-genome

• L123: It is unclear what data set is supposed to be available via SUCEST-FUN and visiting the website didn't help me either – this needs clarifying in both the MS and the website.

Response: Following the previous comment, we declare that data access will be publicly available straight after publication. In the SUCEST-FUN website, a Genome Browse will be available with contigs sequences, tracks for gene Prediction and Annotation, alignment with RNA-seq reads and Sugarcane Assembled Sequences (SAS). We completed the information in the "Availability of data and material" as follow:

L757-764:

"**Availability of data and material**

Genomic data is publicly available at NCBI under GenBank Bioproject PRJNA431722. Contig sequence, gene annotation, alignment with RNA-seq reads and SAS are also available in a genome browser framework at http://sucest-fun.org/cgi-bin/cane_regnet/gbrowse2/gbrowse/microsoft_genome_moleculo_scga7/ or http://sucest-fun.org/wsapp/ (>Cane Genome >Sugarcane Genome Microsoft-Moleculo). The microarray data have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE124990. All data and scripts are also available at GigaDB and in a Github repository (https://github.com/sp80-3280-genome)."

• L124: The word "comprehensiveness" is probably not the right word to use here when you probably mean "completeness"?

Response: We appreciate the reviewer's suggestion but, as presented in the paragraph (L126-136), the indicators suggests that our assembly includes nearly all genes of the species selected as references : 99.4% of sorghum transcripts; 87.5% of ultra-conserved CEGMA; 99.5% of Plantae lineage BUSCO genes; 99.994% and 99.998% of coverage and identity respectively to a *Saccharum* hybrid Chloroplast sequence; 99.850% and 99,927% of coverage and 99,900% and 99,936% of identity respectively of the *S. officinarum* mitochondrial chromosomes 1 and 2, respectively; and 94.4% of SP80-3280 EST database. We have checked the meaning of 'comprehensiveness' and consider it best represents our results.

• L124-134: For the reasons explained previously, I feel these numbers do not accurately reflect the completeness of the gene-space assembly. I need more convincing that all homeologs in the genome are well represented in the assembly; since these numbers could also indicate that only a single homeolog is present in the assembly.

Response: We appreciate the reviewer's comment and would like to clarify that we provide indicators that suggest that our assembly includes nearly all genes from the species selected as references. Concerning homo(eo)logs representation, please also refer below where we provide answers to the next comments.

• L145-146: "A total of 127,940 ESTs (94.9%) have at least one match in the assembly" doesn't really provide confidence that all homeologs in the genome are represented in the assembly.

Response: We have revisited this analysis and included additional evidence (Additional file 1: Fig. S4) which shows that 84.9% of ESTs and 87% of CEGMA show 2 or more matches on the genome, similarly to the results obtained when only single copy genes where used, suggesting that our assembly holds the majority of putative homo(eo)logs in the sugarcane genome. Therefore, we have changed the text as follows:

In the Results section, we have added new data:

L151-156: "Although 10.4% of ESTs (12,966) have a unique hit, what may represent sequencing/assembly issues or genes loss, 84.9% of ESTs (106,133) show 2 up to 15 matches on the genome, reflecting the presence of the majority of putative homo(eo)logs (Additional file 1: Fig. S4A). This result is similar to the search of CEGMA matches against the genome itself using BLASTn. From 235 sequences completely or partially covering CEGMA proteins, 205 has 2 or more hits on the assembly, with most of them (32) with 5 matches (Additional file 1: Fig. S4B)."

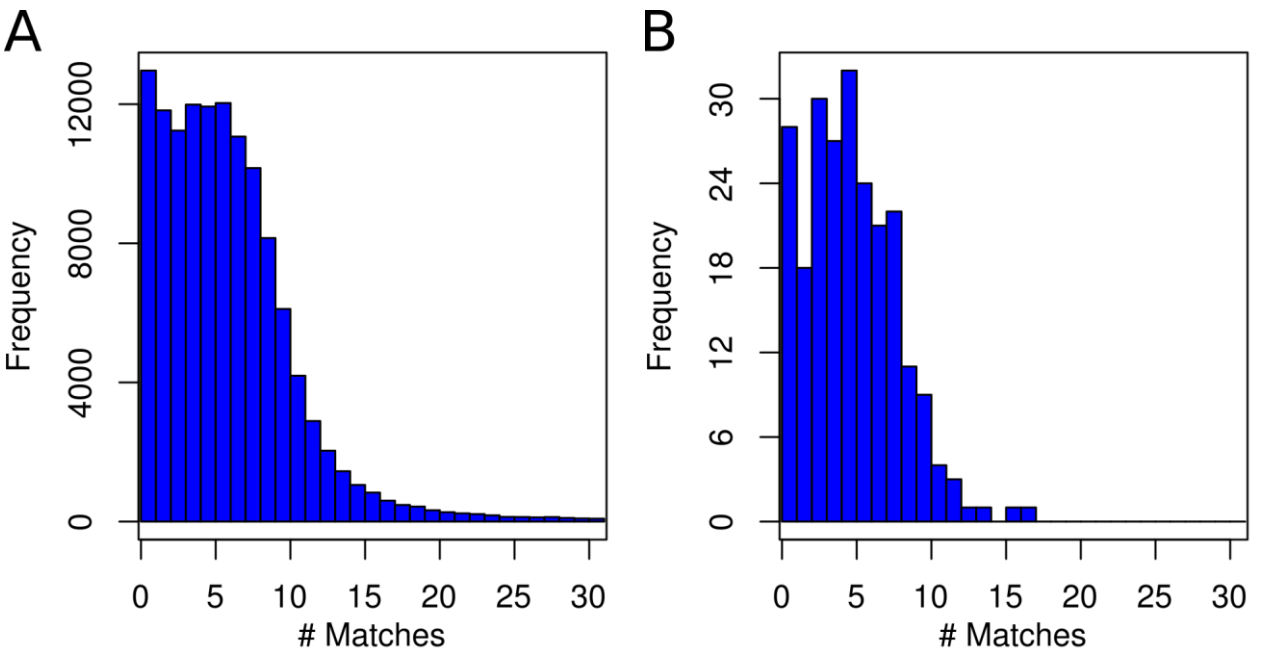


Fig S4: Frequency density of A: Expressed Sequence Tags (ESTs) and B: Core Eukaryotic Genes Mapping Approach (CEGMA) regions alignment on Sugarcane genome assembly. For 127,940

aligned ESTs, 106,133 (84.9%) show 2 up to 15 matches on the genome, while for CEGMA regions, 205 (87.2%) range from 2 to 17 matches on the genome.

In the Methods section, we have added the following:

L498-505:

"**Comparison with Sugarcane ESTs**

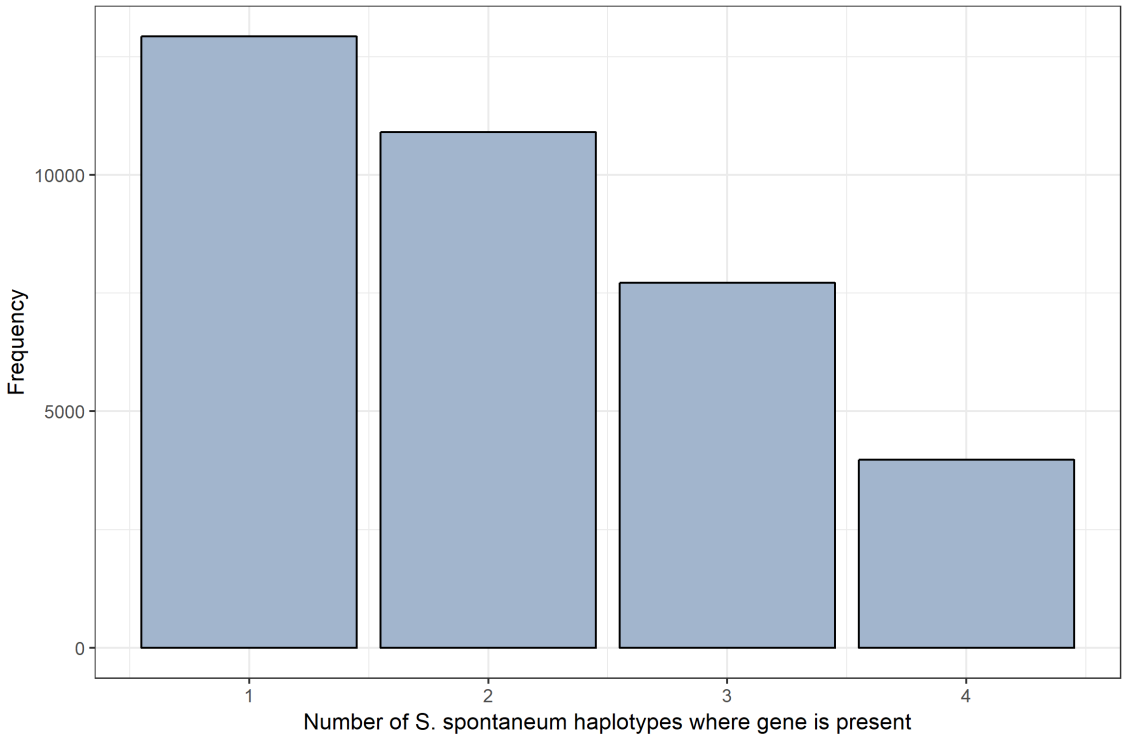
A set of 134,840 ESTs from leaves, internodes and roots samples exclusively from SP80-3280 [20] were aligned to the contigs sequences using SPALN [67] applying mapping and alignment algorithm (-Q 5) and admitting all possible matches for each sequence (-M 1000). Coordinates of aligned ESTs were compared to gene annotation using Bedtools intersect utility [68]. Alignments might be explored through a GBrowse envirioment available at http://sucest-fun.org/cgi-bin/cane_regnet/gbrowse2/gbrowse/microsoft_genome_moleculo_scga7/ or http://sucest-fun.org/wsapp/ > Cane Genome > Sugarcane Genome Microsoft-Moleculo). "

• L149-151: Same comment regarding this statement: "with 93% of the sequences (40,147) matching at least one location in contigs."

Response: We have deleted this sentence since sugarcane assembled sequences (SAS) are derived from the EST data set and produces similar results regarding the overall alignment rate.

• L156-157: This suggests that most genes are present a small number of times. But is this due to some homeologs being lost from the genome or is it an artefact of the genome assembly?

Response: We appreciate the reviewer's comment but there are a few points that must be considered when assessing the number of copies of each gene. Although there are eight or more copies of each chromosome, it is not strictly necessary that each chromosome copy contains all the genes, because of potential gene loss. To that end, we took the table from the *Saccharum spontaneum* (AP85-441) assembly, which is an allele-defined chromosome level genome assembly (available at http://www.life.illinois.edu/ming/downloads/Spontaneum_genome/Saccharum spont alleleTable-Jan 2019.csv) and tabled the presence/absence of each gene in the four assembled sets of chromosomes. The figure below shows that most genes were found in only one or two copies, similar to the results we present in Figure 1A. Indeed, only a minority of genes are present in all four copies. This does not account for gene duplication in the same chromosome, but only shows presence/absence.

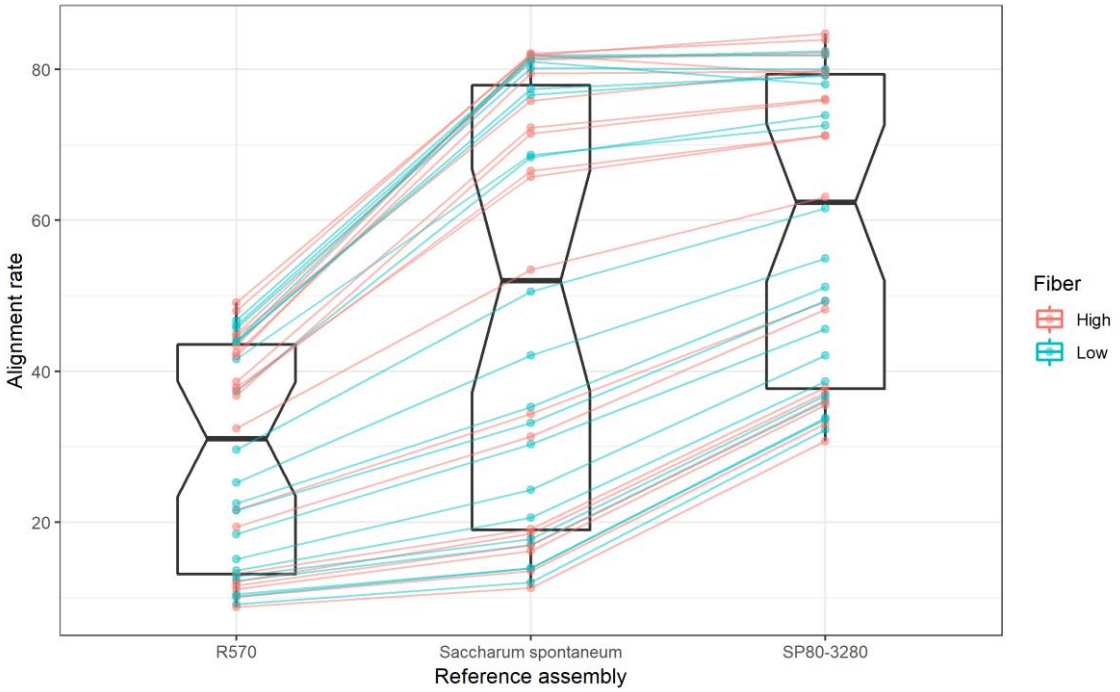


Second, the average depth of coverage with synthetic long reads can be misleading, because of the way they are created. Each long read is in fact a local assembly, based on standard short reads. Repetitive regions are harder to assemble and thus less likely to be represented in the long reads. Thus, the depth of coverage with these long reads is not expected to be uniform. Although the average depth may not be very high, low copy gene-rich regions are expected to be present with higher depth.

Incorrect collapse of homo(eo)logs during the assembly is indeed a possibility. However, it is also possible that (fairly long) identical haplotype blocks exist in the SP80-3280 genome, which would be indistinguishable to the assembly algorithm. Given the fact that some genotypes are used recurrently as parents in sugarcane breeding programs, the pedigree of commercial cultivars often show some level of inbreeding. Genotypes such as Co281, POJ2878, POJ2364 and POJ213 appear multiple times in the SP80-3280 genealogy. With limited chance for recombination, given the small number of generations (crosses) during its breeding process, it is possible that this genome contains IBD (Identity by descent) copies of some chromosome regions.

Finally, incomplete annotation of gene models can result in underestimation of the number of gene copies.

To investigate the value of our assembly as a genomic resource, we used RNA-seq data from Kasirajan et al. (2018) (available at https://www.nature.com/articles/s41598-018-30033-4). These authors created RNA-seq libraries from the top and bottom internodes of 20 different genotypes, including commercial cultivars and introgression lines derived from crosses with wild *S. spontaneum* relatives and *Erianthus*. We used HISAT2 to align these RNA-seq reads against the monoploid R570 hybrid assembly, the AP85-441 tetraploid *S. spontaneum* reference and our SP80-3280 hybrid (results below).



The monoploid R570 assembly is clearly quite incomplete. For the vast majority of samples, our assembly resulted in higher alignment rates than those against *S. spontaneum*. Only in the very high end of alignment rates did the AP85-441 assembly perform better for a few samples - these cases correspond to *S. spontaneum* introgression lines. Libraries with very low alignment rates include those derived from *Erianthus*, as well as some that appear to be problematic (possibly due to unsuccessful library prep). These data show that our genome assembly can be used as a good reference for downstream genomics studies.

• L158: What is the difference between homeologs and "gene copies"? If there is no difference, please be consistent throughout the MS. Perhaps "putative homeolog" might be a suitable alternative?

Response: We appreciate and accept the reviewer's suggestion. We have changed the term "gene copies" to "putative homo(eo)logs" in the manuscript revised version. We only kept "Gene copy count" in the x-axis label of figure 1A to indicate that we used this parameter to estimate sugarcane putative homo(eo)logs from the 1,592 single copy genes from sorghum, rice and *Brachypodium* that matched to the sugarcane genome.

• L165-166: Do you mean antisense transcripts as those being involved in the suppression of translation through mRNA binding or are you merely commenting on the strand on which a gene is encoded? If the latter, I do not understand the relevance/importance of the statements around sense/antisense orientation of genes, especially since most contigs only contain a single gene, the orientation of the contig is irrelevant.

Response: We are referring to antisense transcripts potentially involved in expression regulation. The Trinity version 2.0.6 Sugarcane ORFeome [28] was obtained from full-length first-strand cDNA (FLFS cDNA) synthesis libraries, which maintain strand orientation. Therefore, we can identify antisense transcript, which are largely reported as potential regulators of gene expression and this survey might be useful in future studies.

• L169-170: I am unclear what is being stated here. Do you mean the number of homeologs expressed for a given gene depends on the tissue? Is so, what statistical test shows this is significant?

Response: We appreciate the reviewer's comment and, indeed, we do not provide statistical test to make this statement. Therefore, we have removed this statement from the text.

• L180: ScSuSy is not explicitly defined

Response: We have modified the text as follows:

L178-179: "… phylogenetic analysis of 44 ScSuSy (Sugarcane Sucrose Synthase) CDSs identified …"

• L180: I believe when you mention "members" you mean "operational taxonomic units (OTUs)" in a phylogenetic context?

Response: We have modified the text as follows:

L178-179: "... phylogenetic analysis of 44 ScSuSy (Sugarcane Sucrose Synthase) CDSs identified in the SP80-3280 assembly ..."

• L181: In phylogenetics, the term "clade" has a specific meaning. That is, a group of OTUs which includes an ancestor and all its descendants. To be able to identify all descendants, a phylogenetic tree requires a root to provide evolutionary directionality to the tree. The trees shown do not appear to be rooted, usually by including an outgroup, and as such you should avoid the word "clade". Perhaps use "group" or "grouping"?

Response: We accept the reviewer's suggestion and have changed "clade" to "group" in the revised manuscript.

• L190-191: It is unclear what type of correlation, positive or negative, exists between PAL and lignin. Please be more explicit.

Response: References 38-41 states that silencing PAL genes results on lignin content reduction. Therefore, we accepted the reviewer's suggestion and have modified the text as follows:

L204-205: "Phenylalanine ammonia-lyase (PAL) is the first enzyme in phenylpropanoid biosynthesis [40–42] and silencing its expression has been associated to a reduction in lignin content [40–43]"

• L207: "differentially expressed" implies a statistical analysis of differential gene expression using limma, edgeR, DESeq or similar. If this is not the case please reword.

Response: We apologize for this, we meant evidence of expression (FPKM>1) and have modified the text as follows:

L233-235: "The sheer number of sugarcane genes found so far, the large size of multi-gene families and the evidence that not all homo(eo)logs are expressed point to a very complex role of regulation in the determination of phenotypic differences."

• L224: Table S8c does not exist.

Response: We have modified the text as follows:

Line251: "… Table S8 and …"

• L234: This is the first time referencing "two modern cultivars" so it would be wise to either explicitly state which these are or simply refer to them by name.

Response: We have modified the text as follows, to emphasize that SP80-3280 and R570 are two modern cultivars:

L261: "The two modern cultivars (SP80-3280 and R570) have fewer TE counts"

• L254: "Susy" Should this be "SuSy"? If so, please correct here and in the supplementary files.

Response: Yes. We apologize for this mistake and have modified the text in both MS and Supplementary files.

• L257: Fig. 4 shows the distribution of SNP's relative to the sorghum genome. However, the authors make no reference or discuss the patterns of distribution between or within chromosomes. Without such an analysis and discussion, it would make sense to report the SNP numbers in a table.

Response: We accept the reviewer's suggestion and have added a sentence regarding chromosome SNV distribution, as follows:

L291-294: "The preservation of as many as four functionally different alleles at a locus, with cases observed on all except one chromosome (Chr 10 - **Fig. 4**), is consistent with the well-known heterozygosity of sugarcane cultivars and associated susceptibility to inbreeding depression."

• L258-259: I am unaware of what "premature splicing" and "translation initiation" SNVs are. Please clarify. Please use of a variant effect prediction tool such as SNPeff or VEP, for determining the effects of SNPs on CDS'.

Response: We appreciate the reviewer comment and apologize for this. There is a missing comma in "premature splicing". We have clarified the terms by changing "premature" to "premature termination", "splicing" to "erroneous splicing" and "translation initiation" to "incorrect translation initiation" in the manuscript (L297-298 below), Figure S8 and Table S11.

SNVs were called using samtools v1.1 and bcftools v1.1 and large-effect SNVs were identified as those mapped to coding regions, splicing sites, stop codons and transcription initiation sites.

L297-298: "Further, 1,334 SNVs that differentiate sugarcane from sorghum in 585 single copy genes include frameshifts, premature termination, erroneous splicing, loss of stop codons and incorrect translation initiation…"

• L259: Please provide S11 as a VCF file, the defacto standard for reporting variants and can also include the variant effect annotations from tools such as SNPeff, VEP etc.

Response: We appreciate the reviewer's comment and declare that all data (including VCF file 'SNV_annotation.vcf') was deposited at GigaDB at the time of submission.

• L271: Please include the percentage (17.6%) of all contigs with at least two predicted genes

Response: We have included the percentage, as suggested.

L311: "… 79,094 (17.6%) contigs …"

• L275: Please include the percentage (2.1%) of all contigs with at least two synteny anchors

Response: We have included the percentage, as suggested.

L315: "… 9,319 (2.1%) SP80-3280 contigs …"

• L276: Please include the percentage (1.8%) of all contigs that were fully syntenic

Response: We have included the percentage, as suggested.

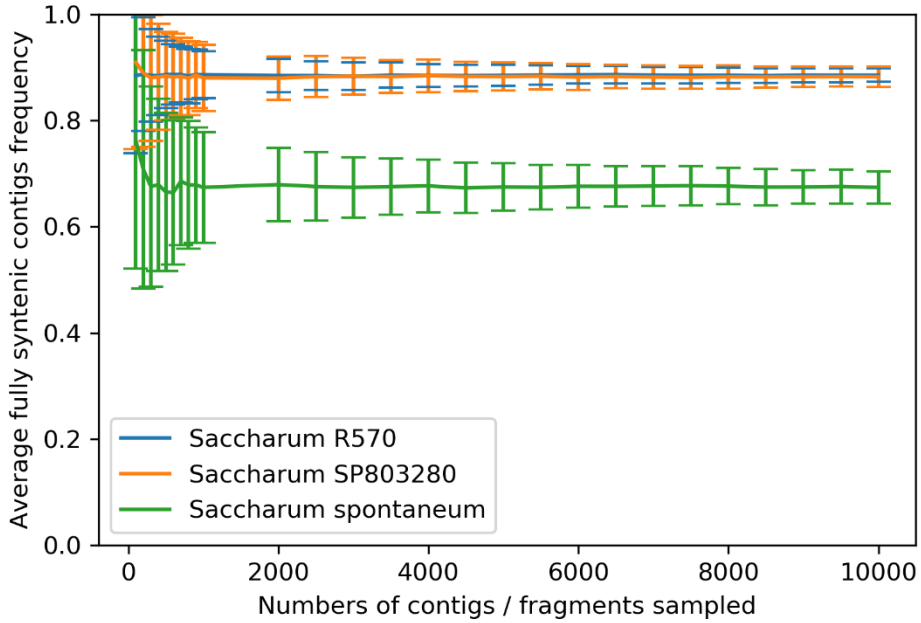
L316: "… 85% (7,906 – 1.8% of all contigs) of these contigs were fully syntenic …"

• L284-286: Given the above percentages I find it difficult to accept such strong assertions as "our assembly is enriched in genomic neighborhoods that are co-linear to sorghum" and "our results agree with widespread findings on the conservation of gene order among grass species" since this is based on only a fraction (1.8%) of all contigs in the assembly.

Response: We used the term enriched to indicate that, from the contigs with two or more marker genes, the majority (85%) was co-linear with Sorghum. But we agree with the reviewer that the use of "our assembly is enriched" was not correct. On the other hand, although our sample of contigs with two or more marker genes is small, it captures the overall pattern of gene order conservation between SP80-3280 and *Sorghum bicolor*.

Complementing the argument presented in the manuscript (L317-330), we address the minimal sampling size and concluded that only 0.4% of contigs would be sufficient. We used Monte Carlo

simulations with varying numbers of sampled contig fragments, from 100 to 10,000 and results are presented in the figure below.



We show that our measure of co-linearity, the proportion of real or simulated contigs where all marker genes are co-linear with Sorghum orthologs, doesn't change when simulations are performed with more than 2,000 contigs. In conclusion, we have rewritten the sentence to more accurately reflect that our data only suggests similar levels of short-range gene order conservation among the genomes analyzed. We have also rewritten the methods section to better explain and emphasize how our analysis was performed, as follows:

In the results section:

L318-330: "To evaluate the effect of SP80-3280 assembly fragmentation on the number of segments with conserved gene order ("syntenic blocks") per contig, we used a Monte Carlo method to simulate the fragmentation of the chromosomes and contigs of the *Saccharum* R570 and *S. spontaneum* genomes. We performed 1,000 rounds of simulation for each genome and, at each round, sampled 10,000 random fragments from each of these two genomes, while simultaneously sampling the same number of contigs from SP80-3280's assembly. Sampled contigs and contig fragments were constrained to follow the distribution of the number of genes per contig observed for the full SP80-3280 assembly. The number of syntenic blocks on each fragment was then evaluated and the relative frequency of contigs/fragments per number of syntenic blocks is shown in additional file 1, Fig. S10C. We observed that contigs and fragments harboring a single syntenic block are sampled at similar frequencies in all genomes analyzed. While an increase in sequencing coverage would lead to improved estimates of co-linearity, our analysis of the small subset of contigs with two or more marker genes suggests that levels of genomic rearrangement in SP80-3280 are similar to those expected anywhere in the genomes of the other two *Saccharum* species."

In the methods section:

L700-708: “In order to evaluate the effect of genome fragmentation on our estimates of gene conservation, a Monte Carlo simulation of chromosome fragmentation was performed on the R570 and *S. spontaneum* genomes. We sampled 10,000 random regions of the R570 and *S. spontaneum* genomes, with fragment lengths constrained to follow the distribution of contig lengths observed for SP80-3280. We performed 1,000 rounds of these simulated fragmentations, every time allowing genomic fragments (and the genes within them) to be chosen randomly throughout the genome, with no bias to marker genes. We accessed the degree of conservation through the fraction of contigs with two or more marker genes that were found in the same order in the *Saccharum* genome fragments and in the *S. bicolor* genome.”

• L287-288: The assertion that “our assembly does not contain an excess of chromosomal rearrangements, as would be expected if there was a significant amount of chimeric contigs.” is based more on a lack of evidence for chimeric contigs rather than evidence for no chimeric contigs since most contigs contain <= 1 gene for synteny and chimeric detection.

Response: We respectfully disagree with the reviewer that our conclusions are based more on the lack of evidence for chimeric contigs. In fact, it is based on the observation that levels of genomic rearrangement detected in the small fraction of SP80-3280 contigs with two or more marker genes are similar to those observed for random fragments of the other two Saccharum genomes, despite the fact that this is a completely independent assembly. This observation is also reinforced by comparison to random fragments spanning all the length of the contigs of the other two species. To make this point clearer and properly align our inferences with our data, we have changed this sentence as follows:

L328-330: “… our analysis of the small subset of contigs with two or more marker genes suggests that levels of genomic rearrangement in SP80-3280 are similar to those expected anywhere in the genomes of the other two *Saccharum* species.”

• L306-307: It is unclear if the authors are referring to 12.25% of the genome sequence or something else. Please reword to remove ambiguity.

Response: As stated in Zhang et al. 2018 [14]: "In the modern hybrid sugarcane SP80-3280, approximately 12.25% of sequences are contributed by *S. spontaneum*." Therefore, we are indeed referring to genome sequence. We accept the reviewer's suggestion and have modified the text as follows:

L348-349: “Approximately 12.25% of the SP80-3280 genome sequence is of *S. spontaneum* origin [14], …”

• L308-309: I am not convinced by the results shown that the assembly does indeed adequately represent all the homeologs present in the genome of SP80-3280 for the reasons stated above.

Response: As explained before, we provide indicators, which suggests that our assembly includes nearly all genes from the species used as references and from 1 to 15 homeologs. We

do not state that we have represented all the homeologs in the genome. That is why we use the word “comprehensiveness” and not “complete”.

• L312-313: Please expand on the reasons for homeologs only being “present in up to 15 copies in the SP80-3280 assembly”. Discussion points should include homeologs missing from the assembly due to sequencing/assembly issues as well as loss of genes in the genome itself.

Response: We accepted the reviewer’s suggestion and have modified the text as follows:

L355-360: “Although for sugarcane modern variates we expect eight or more copies of each chromosome, it is possible that each homolog does not contains a copy of every gene, because of potential gene loss. In addition, it is also possible that some homeologs were not identified in our assembly because of assembly or sequencing difficulties in regions with highly repetitive sequences. Thus, for sugarcane genes found to correspond to single-copy genes in diploid grasses, or that matched to CEGMA genes or that had an EST correspondent, we found mostly from 2 to 15 copies in the SP80-3280 assembly …”

• L316: The data presented does not allow a causal relationship to be determined between the upstream sequence variation and expression patterns. Therefore, “causes” should not be used.

Response: We have modified the text as follows:

L363: “The differences in gene upstream sequences may potentially affect the expression level …”

• L320: Again, I’m not entirely convinced the data presented does indeed show that “our assembly which discriminates homo(eo)logs”

Response: We appreciate the reviewer’s comment. However, as we addressed in the previous comments, we provide evidence that our assembly may be indeed discriminating homo(eo)logs. Therefore, we would like to keep the statement.

• L326: Change “might by” to “might be”

Response: The sentence including this mistake was deleted due to restructuring of the text.

• L343: Drop “On the other hand,”

Response: Same as previous comment.

• L362-363: I do not agree with the assertion that "4 out of 6 clusters correspond to single chromosomes" since most show almost half of each cluster linking to a variety of other chromosomes. It would be beneficial to discuss this discrepancy in the MS

Response: The clusters of contigs identified are certainly not individual chromosomes. We found a consistent pattern when comparing SP80-3280 to *S. spontaneum* and R570 when both genomic and protein sequences were used. Therefore, it may indicate a potential advance towards a chromosome level assembly. To address this issue, we have modified the text as follows:

L375-378: "The majority of proteins predicted from chromosomes 1, 2, 3 and 4 (in both *S. spontaneum* and R570) have their best matches located in SP80-3280 contigs from clusters 2, 5, 6 and 1, respectively (Fig. 5B and C). On the other hand, clusters 3 and 4, which contain contigs matching to multiple chromosomes, including those in which chromosomal rearrangement events were demonstrated …"

• L369-370: This sentence seems a little incomplete/clunky so please reword.

Response: We have modified the text as follows:

L381-384: "Assembling the genome of a polyploid interspecific hybrid is of especially high value for breeders. The assembly, gene prediction, and annotation provided can bridge long standing gaps of knowledge allowing them a more efficient use of genomic tools. Sugarcane's large autopolyploid genome, predominant clonal propagation, and need for extensive phenotyping to determine breeding values ..."

• L381: Again, I'm not convinced that the gene-space is "resolved" in terms of all the homeologs being present. I also think it would be beneficial to use the same adjective when describing the "resolved", "completeness", "comprehensiveness" of the assembly.

Response: We have changed the word "resolved" for "presented" as follows:

L394-396: "The presented gene-space of the sugarcane genome is a fundamental step towards a high-quality chromosome resolved assembly from a current commercial hybrid."

• L285-386: As stated before, I think that the limited data on which measures of synteny could be derived is lacking. As such, the assertion that the assembly is "highly syntenic" is too strong.

Response: We have removed the statement and modified the text as follows:

L396-398: "The genome sequence released for this interspecific polyploid supports its recent allotetraploid nature, reveals differences in promoter regions associated to a diverse gene expression pattern and transposable elements contributing to fine tuning of the sugarcane genome."

• L398: The authors should address any issues with the type of library prep that could give rise to chimeric contigs, especially in the context of a large, polyploid species such as sugarcane which also has a high proportion of repetitive elements. For instance, it is my understanding that each 384-well plate will contain multiple genomic fragments and that these will all get the same barcode/index. As such, if these genomic fragments contain similar enough sequences, chimeric contigs could easily be produced.

Response: According to Illumina's Long Reads Pipeline User Guide, the number of fragments in each well is relatively low, which facilitates the assembly process as there are fewer repetitive sequences in the input data to confound the assembly. The fragments in each one of the 384 wells in the plate are fragmented and barcoded with unique indices. In addition, the haploid nature of the input fragments eliminates the need to accommodate heterozygous variants and thus allows for more aggressive separation of repeat copies. To clarify this issue, we have modified the text as follows:

L414-416: "The fragments from each well were amplified, fragmented and barcoded with unique indices to create 26 TruSeq Synthetic Long-Read DNA libraries."

• L415: "we modified the quality scores" - please describe how they were modified

Response: In our manuscript, we wrote, "Since synthetic long reads are very accurate and some of the base qualities exceeded this upper bound, we modified the quality scores of our long-read data to allow them to be appropriately parsed." That means we modified the quality score in FASTQ file to detour CA (Celera Assembler) quality score check. We'll update the manuscript accordingly, as follows:

L427-429: "Since synthetic long reads are very accurate and some of the base qualities exceeded this upper bound, we transformed the quality scores of our long-read data (FASTQ file) to allow them to be appropriately parsed."

• L417: "incurred" is the wrong word, perhaps "occurred"?

Response: We have modified the text and used "resulted", as follows:

L430-431: "The substantial number of contigs generated initially (roughly 450,000, half of them singletons) resulted in several files in a folder that hindered I/O operations."

• L419: Correct capitalisation of FastA to FASTA

Response: We have corrected this issue in the revised version, as follows:

L432: "… contained not more than a thousand contig FASTA files, …"

• L413-422: Please provide an online code repository containing the modification made to CA. These are required to reproduce your results and may help others will similar issues.

Response: As stated before, we modified only the quality score in FASTQ file to detour CA (Celera Assembler) quality score check.

• L425: Correct capitalisation of PACBio to PacBio.

Response: We have removed this from the manuscript.

• L433: Please provide parameters for BWA-mem

Response: We have modified the text as follows:

L445: "… with BWA mem, using default parameters."

• L440-444: The level of detail here is insufficient to fully understand how this comparison was performed and how hits were filtered. The aim of such a comparison is to identify CDS' against the genomic contigs. An alignment with BLASTn will generate multiple HSPs (e.g. 1 HSP per exon) and thus will generate a non-contiguous alignment representing the whole CDS. How have the authors constrained the alignments further to avoid spurious hits?

Response: We appreciate this comment and understand the concern. For any CDS with multiple HSPs against the same contig that passed the filtering criteria, we used the union of such hits, excluding any potential overlap. Given that most contigs contained only one or two genes, we expect very little influence of spurious hits to different gene regions.

• L454: "parted" is the wrong word, perhaps "partitioned"?

Response: We appreciate this comment and have modified the text as follows:

L470: "… we partitioned sugarcane contigs ..."

• L454: "volume" is the wrong word, perhaps "number" or "length" would be more precise?

Response: We appreciate this comment and have modified the text as follows:

L470: "… into six groups with similar length and ..."

• L459-472: The main aim of this method is to identify contigs assembled from reads of chloroplast or mitochondrial genome origin. How have the authors controlled for contigs which harbour highly conserved genes (e.g. rRNA) which can be encoded in both nuclear and cytoplsmic compartments and result in false positive identification of mitochondrial/chloroplast contigs? Especially given the low identity threshold of 70%. Perhaps the use of read depth could facilitate this as chloroplast and mitochondrial reads would be expected to be in higher abundance?

Response: We appreciate the reviewer's comments, but we do not expect chimera at the rDNA operon between nuclear and organellar genomes as they belong to different domains of life (Bacteria and Eukaryotes). The nuclear ribosomal operon is larger in size and its origin is eukaryotic, while the chloroplast and mitochondrial operons are smaller due to bacterial

ancestry. However, potential other specific gene transfer from organellar DNA to the nucleus cannot be ruled out until completion of SP80-3280 nuclear chromosomes. Nevertheless, the reconstructed organellar genomes presented has no missing genes; only mismatches and small gaps. We have aligned our general assembly to Saccharum reference organellar chromosomes, one chloroplast (NC_005878.2, from SP80-3280) and two mitochondrial genomes (LC107874.1 and LC107875.1, from *S. officinarum*). We identified 51,768, 2,482 and 909 contigs (mean size of 4Kb and overage depth higher than 20x) mapping to NC_005878.2, LC107874.1 and LC107875.1, respectively. In addition, reconstructed organellar genomes were 99.998%, 99.900% and 99.936% identical to NC_005878.2, LC107874.1 and LC107875.1, respectively. Since our DNA samples were prepared from leaf tissue, this supports the higher plastid genome incidence. Organellar genome references were retrieved from the NCBI and correspond to the Chloroplast genome of *Saccharum* Hybrid (SP80-3280) and Mitochondrial chromosomes from *S. officinarum*. In support to use these organellar genomes as reference, previous works has demonstrated that the female parent of *S. officinarum* contributed about 80% of the chromosome to the genome of *Saccharum* Hybrid (D'Hont A., 2005, Cytogenet Genome Res. 2005; 109(1-3):27-33).

• L468-469: Is it correct that the reference chloroplast genome (NC_005878.2) has 100% of it's bases represented by the assembled contigs and that these are 100% identical, no gaps and no additional bases?

Response: We appreciate the reviewer comment and apologize for this. The actual value is 99.998%. The chloroplast genome reference from NCBI (NC_005878.2) is originated from the same cultivar that we have sequenced and assembled (SP80-3280), thus we expected to find high similarity for the reconstructed chloroplast genome. Therefore, we altered the manuscript and added more details about the comparison between the reconstructed chloroplast genome and the NCBI reference, as follows:

L482-483: "The contigs used for genomes reconstruction presented mean size of 4Kb, with coverage depth higher than 20x."

L485-488: "The reconstructed consensus sequence aligned against the chloroplast genome presented 99.994% and 99.998% of coverage and identity respectively, and there were identified only 6 mismatches and 2 gaps, most of them located in intergenic regions and in one of the rRNA23S copies with protein frame preservation."

• L469-470: Similarly, a similar question for the two reference mitochondrial genomes? How similar then are the two reference mitochondrial genomes to each other?

Response: The two reference mitochondrial genomes (LC107874.1 and LC107875.1) from NCBI are almost completely different. There were found similarities only in the genomic regions corresponding to the nad2, nad5, trnL, trnM and trnS genes. Similarly to previous comment, we have added more details in the paper for the reconstructed mitochondrial chromosomes and comparisons with the NCBI *S. officinarum* reference chromosome 1 (LC107874.1) and chromosome 2 (LC107875.1).

L489-496: "The alignment against mitochondrial chromosomes 1 and 2 presented 99.850% and 99,927% of coverage and 99,900% and 99,936% of identity, respectively. The consensus

sequences were annotated using their respective NCBI references with the CLC tool "Annotate from Reference", where all genes, tRNAs, rRNAs and miscellaneous features were totally transferred. For the mitochondrial chromosome 1, 237 mismatches and 63 gaps were identified, most of them present in intergenic regions and only 2 mismatches in 2 rRNA genes, with proteins frame preservation. And for chromosome 2, we identified a region composed by 19 N's inside a repetitive AT's region. In addition, the reconstructed chromosome has 57 mismatches and 16 gaps, all of them present in intergenic regions."

• L506-521: See L158 comment with regards to "gene copies"

Response: We have accepted the reviewer's suggestion and modified the title as follows:

L540: "Identification of Putative Homo(eo)logs and Count Estimation"

• L511: Is the "80% nucleotide identity" the same as the "80% identity" stated on L513? If so, then there is some repetition to remove.

Response: We thank you for pointing this out and have corrected the text accordingly:

L545-546: "We filtered alignments with at least 80% nucleotide identity, based on Wang *et al*. [50], covering at least 70% of both the sugarcane and sorghum sequences."

• L518: What is the rationale for discarding clusters with > 16 sugarcane genes? What was the distribution of cluster sizes prior to filtering? I am curious since plots presented show only up to 12 putative homeologs rather than 16.

Response: We thank you for bringing this to our attention. The distribution of cluster sizes corresponds to that of gene copy counts, shown in Figure 1A. When getting estimates of sequence differentiation, our goal was to focus on putative alleles (homo(eo)logs), while excluding any possible paralogs. To that end, we started by using genes that are present in a single copy in the genomes of diploid grasses, and thus likely to be present in single copy in sugarcane as well. We noted that a single cluster had 21 gene copies, which is higher than the number of chromosome copies expected for sugarcane. For this reason, we decided to remove this cluster from this analysis. We have now revised the text to reflect this motivation:

L550-552: "We aligned the coding sequences for each pairwise combination in each gene cluster, using BLAT v35 [80] (*–minIdentity=0 –minScore=60*). One of the clusters had 21 putative homo(eo)logs, which is higher than the number of chromosome copies expected for sugarcane and was discarded from the analysis."

• L523: See L158 comment with regards to "gene copy"

Response: We have accepted the reviewer's suggestion and modified the title as follows:

L558-559: "… of the predicted sugarcane putative homo(eo)logs."

• L530-531: This seems ambiguous. Do you mean that for contigs which do not contain the required length of upstream sequence (due to gene being close to the end of the contig) were excluded? How many such sequences were excluded? Is this the reason for the incomplete lines in Fig 3C and 3D?

Response: In this section, we describe the methods used to build Figure 1B, in which we show the dissimilarity level, considering both mismatches and gaps (which may alter the initial sequence length – 100pb, 500pb and 1000pb). Therefore, we only considered alignments up to 20% shorter or longer than the expected sequence length.

• L534: See L158 comment with regards to "gene copies"

Response: We have accepted the reviewer's suggestion and modified the title as follows:

L567: "To investigate the occurrence of frameshift mutations between putative homo(eo)logs, …"

• L550: Correct capitalisation of "perl" to "Perl"

Response: We have corrected this issue in the revised version.

L583: "An in house Perl script was used …"

• L552: Correct capitalisation of "perl" to "Perl"

Response: We have corrected this issue in the revised version.

L585: "We developed in-house Perl and R language …"

• L556-558: Same comment as for L165-166

Response: As stated previously, The Trinity version 2.0.6 Sugarcane ORFeome [28] was obtained from full-length first-strand cDNA (FLFS cDNA) synthesis libraries, which maintain strand orientation.

• L566: Change "Based in" to "Based on"

Response: We have corrected this issue as follows:

L599: "Based on BLAST …"

• L566: Please state what keywords were used so that other may reproduce the results

Response: We have modified the text as follows:

L599-600: "Based on BLAST and keyword search ('Phenylalanine ammonia-lyase', 'PAL' and 'EC:4.3.1.24') in two databases ..."

• L569: States that BLASTn was used to identify PAL genes in the predicted proteins. This is not possible since BLASTn performs a search of a nucleotide sequence against a database of nucleotide sequences, not protein sequences. It is most likely either BLASTx or BLASTpplease clarify.

Response: We apologize for this mistake and have corrected the text as follows:

L602: "… genes for a BLASTx …"

• L576: The use of "use all sites" is sensitive to gaps in the alignment. Please comment on the quality of the alignments, especially in terms of gaps and perhaps even provide a supplemental containing the MSA for readers to visualise.

Response: Since we are comparing sequences from few species (*Sorghum bicolor*, *Saccharum spontaneum* and sugarcane hybrids – SP80-3280 and R570), insertions/deletions (gaps in the alignment) are also important polymorphic source of information. Therefore, we decided to use all sites for gene tree reconstruction. In addition, we reinforce that all data was deposited at GigaDB at the time of submission, including the 'PAL_clustalw.meg' and 'Susy_clustalW.meg' files, which contains the alignments results for visualization.

• L639: I think instead of "reads" you probably mean "contigs"?

Response: We apologize for this and have corrected the text as follows:

L675: "… and contigs with mapping quality larger than 20 …"

• L639: When you say "alignment score larger than 20" do you mean the "MAPQ value"?

Response: Yes. We have corrected the text as follows:

L675: "… and contigs with mapping quality larger than 20 …"

• L640: Correct capitalisation of "python" to "Python"

Response: We have corrected this issue as follows:

L676: "Using in-house Python scripts, …"

• L644-645: Please provide details on how the large effect SNVs were identified

Response: We have added the text as follows:

L680-681: "SNVs mapping to coding regions, splicing sites, stop codons and transcription initiation sites were classified as potential large-effect SNVs."

• L681: Consistent capitalisation of “BlastP” to “BLASTp”?

Response: We have corrected this issue in the revised version.

L685: “… similarity inferred from best BLASTp …”

• L716: Please ensure that accession GSE124990 is made public as it is still private until Dec 21, 2021

Response: As stated before, we declare that data access will be publicly available straight after publication.

Table 1

• Remove the “Mean read length” row from the “sequence and assembly data” part of the table, and merge the info into the row above. This will ensure consistency with the “Contigs Length” row.

Response: We have modified the text in Table 1, as suggested.

• Change “Genome” row label to “Genome Coverage”

Response: We have modified the text in Table 1, as suggested.

• Change “6,6 Gb” in the “Total Sequence” row to “6.6 Gb”

Response: We have modified the text in table 1, as suggested.

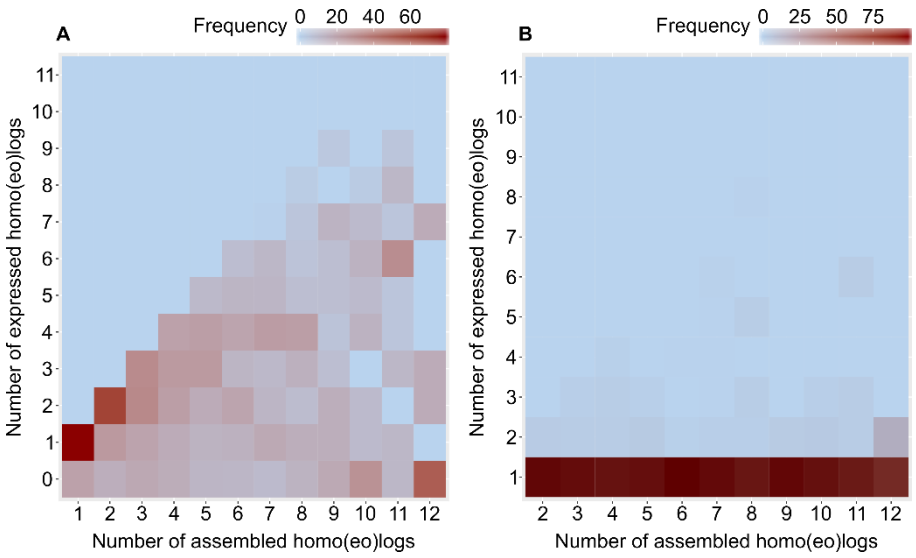
• Change “450.609” to “450,609” in the “Number of unitigs” row

Response: We have modified the text in table 1, as suggested.

Figure 2

• I think the two plots would be more informative, if the “heat” was scaled/expressed as a percentage of the number of genes with a given number of homeologs. i.e. the sum of each column. That way, the colours for the genes with a higher number of homeologs (A) and higher number of expressed homeologs (B) would not wash out into the background.

Response: We have accepted the reviewer’s suggestion and have modified the “heat” scale in Figure 2.



• Add “assembled” to the x-axis labels. i.e. “Number of assembled homo(eo)logs”.

Response: We have modified the x-axis labels in Figure 2, as suggested.

Figure 3

• Legend: Please add additional information to make the figure more “standalone”. e.g. the type of phylogenetic analysis, how the branch support values are calculated, why some branch support values are not shown and better orientation to quite an information rich figure.

Response: We have modified the Figure 3 legend, as suggested.

L1067-1081: “**Fig. 3 – Phylogeny, putative regulatory regions and expression of sucrose synthase (SuSy) and phenylalanine-ammonia lyase (PAL) gene family.** Phylogenetic analysis of (**A**) SuSy and (**B**) PAL genes from SP80-3280, R570, *S. spontaneum*, and sorghum. SuSy sequences from *Saccharum* ssp [34] were also included. For both SuSy and PAL, nucleotide sequences (CDS) were aligned with CLUSTALW [90] software in MEGA 7.0 [91] and maximum likelihood trees were constructed with default parameters except for 1,000 bootstraps (only bootstraps values over 20 are shown). Core promoter analysis (gray columns in C and D) using TSSPlant [93] suggests ScSuSy2 (**C**) and most ScPAL (**D**) as TATA-less (absence of black squares). Transcription factor binding sites (TFBS) prediction (colored symbols in **C** and **D**) using MEME [94] and MotifSampler [95] suggest specific motif for each group (ScSuSy1, ScSuSy2, ScSuSy5 and PAL I, PAL III, PAL Va and PAL Vb). The three SP80-3280 PAL genes marked (* in **D**) are present in the same contig. Transposable elements (TEs) were identified within 10 kb upstream from the gene (**C** and **D**). Heatmap analysis of RNA-Seq data [28] (expression profile in **C** and **D**) shows more pronounced expression in SP80-3280 internodes (I1 and I5) of ScSuSy1, ScSuSy2, ScSuSy5 and PAL from group V. RNA-Seq of leaf tissues (L) indicates more pronounced expression of ScPAL from groups II and III. ScSuSy3 presents high numbers of TFBS and TE and low expression in all samples.”

Figure 5

• It seems to me that Clusters 1, 2, 5 and 6 have correspondence to more than a single chromosome from both S. spontaneum and R570. See also comment for L362-363

Response: As mentioned before, the clusters of contigs identified are certainly not individual chromosomes. We found a consistent pattern when comparing SP80-3280 to *S. spontaneum* and R570 and it may indicate a potential advance towards a chromosome level assembly. (See above where we edited to L375-378)

Other points to address:

• Ensure consistent use of a comma and full-stops, in numbers, for a thousand separator and decimal points throughout the MS and figures, including supplementary files. e.g.:

◦ Fig. S1

Response: We have modified the text, as suggested. In figure S1, the x-axis indicates percentage from 0.00 to 100.00%.

◦ L215: missing comma from “3280 plants”

Response: In this sentence, the “3280” is part of the sugarcane genotype’s name: “… gene expression data of SP80-3280 plants grown in field conditions…”

◦ L355: comma used instead of decimal point for “0,97%”

Response: We have corrected the text, as suggested.

◦ L417: “450.000” a full-spot is used instead of a comma for the thousand separator

Response: We have corrected the text, as suggested.

◦ L594: Add thousand separator “,” to “1500 nt”

Response: We have corrected the text, as suggested.

◦ See comments above regarding table 1

Response: We have corrected table 1, as suggested.

• When presenting numbers with exponents, please use proper scientific notation rather than using E-notation, which is generally used by displays incapable of superscript (e.g. calculators)

◦ L463: Instead of 1.0E-15 use 1x10-15

◦ L491: Instead of 1e-5 use 1x10-5

◦ L511: Instead of 1e-6 use 1x10-6

◦ L581: Instead of 1e-6 use 1x10-6

◦ L588: Instead of 1e-3 use 1x10-3

◦ L681: Instead of 1e-5 use 1x10-5

Response: We have corrected the text, as suggested.