Check for updates

DATA NOTE

Initial genome sequencing of the sugarcane CP 96-1252 complex hybrid [version 1; referees: 2 approved]

Jason R. Miller [1], Kari A. Dilley[1], Derek M. Harkins[1], Manolito G. Torralba[2], Kelvin J. Moncera[2], Karen Beeri[2], Karrie Goglin[2], Timothy B. Stockwell[1], Granger G. Sutton[1], Reed S. Shabman[1]

[1]J. Craig Venter Institute, Rockville, MD, 20850, USA
[2]J. Craig Venter Institute, La Jolla, CA, 92037, USA

v1 **First published:** 17 May 2017, **6**:688 (doi: 10.12688/f1000research.11629.1)
Latest published: 17 May 2017, **6**:688 (doi: 10.12688/f1000research.11629.1)

Abstract

The CP 96-1252 cultivar of sugarcane is a complex hybrid of commercial importance. DNA was extracted from lab-grown leaf tissue and sequenced. The raw Illumina DNA sequencing results provide 101 Gbp of genome sequence reads. The dataset is available from https://www.ncbi.nlm.nih.gov/bioproject/PRJNA345486/.

This article is included in the Global Open Data for Agriculture and Nutrition gateway.

This article is included in the Data: Use and Reuse collection.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
version 1 published 17 May 2017	✓ report	✓ report

1 **Paulo Arruda** , State University of Campinas (UNICAMP), Brazil

2 **Jeremy R. Shearman** , National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand

Discuss this article

Comments (1)

Corresponding author: Jason R. Miller (jmiller@jcvi.org)

Competing interests: No competing interests were disclosed.

How to cite this article: Miller JR, Dilley KA, Harkins DM *et al.* **Initial genome sequencing of the sugarcane CP 96-1252 complex hybrid [version 1; referees: 2 approved]** *F1000Research* 2017, **6**:688 (doi: 10.12688/f1000research.11629.1)

Copyright: © 2017 Miller JR *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Grant information: This work was funded by US Department of Homeland Security (contract HSHQDC-15-C-B0059).

First published: 17 May 2017, **6**:688 (doi: 10.12688/f1000research.11629.1)

Introduction

Sugarcane is an important crop for food and energy production. The genomes of modern cultivars are hybrids of species that are themselves polyploid; see for example (Vilela *et al.*, 2017). Selected genomic BAC sequences have been sequenced and assembled (de Setta *et al.*, 2014) (Okura *et al.*, 2016). Chloroplast and mitochondrial genomes have been published (Asano *et al.*, 2004) (Shearman *et al.*, 2016), as have several transcriptomes (Cardoso-Silva *et al.*, 2014). Whole genome sequence assemblies have not been published. CP 96-1252 is the top commercial sugarcane cultivar in Florida, USA (Sandhu & Davidson, 2016). CP 96-1252 was developed by USDA-ARS, the University of Florida, and the Florida Sugar Cane League and released to growers in 2003. CP 96-1252 is a complex hybrid of *Saccharum officinarum* L., *S. barberi* Jeswiet, *S. spontaneum* L., and *S. sinense* Roxb. amend. Jeswiet (Edmé *et al.*, 2005). Toward better understanding of this cultivar through its genome sequence, DNA reads were generated and made public.

Methods

Using lab-grown plantlets, kindly provided by USDA, 14 g of tissue was harvested from the leaves of *Saccharum* hybrid cultivar CP 96-1252 (Reg. no CV-120, PI 634935, NCBI taxon ID 1983727). DNA was extracted from purified plant nuclei at Amplicon Express (Pullman, WA, USA). Separately, DNA was extracted from whole cells at JCVI (Rockville, MD, USA) using a Qiagen Plant DNA isolation kit. Extracted DNA was fragmented and size selected on the Blue Pippin (Sage Scientific) prior to library construction to ensure a 260 bp insert size. Standard Illumina PE libraries were generated using the NEBNext kit (NEB). Libraries were size selected, QC'd and quantified by qPCR prior to sequencing. Barcode BS78 AGCCATGC was used for the nuclei prep library and barcode BS79 AGGCTAAC was used for the cell prep library. The libraries were generated and sequenced at the JCVI sequencing core in La Jolla, CA, USA. To test for bacterial contamination, both DNA samples plus negative controls were used to generate amplicon libraries targeting the V4 16S region followed by Illumina MiSeq sequencing. These reads were processed by a pipeline using usearch version 8.1.1.1861 for clustering (Edgar, 2017), mothur version 1.36.1 for taxonomic classification (Schloss *et al.*, 2011), and the SILVA SSURef NR99 123 database for reference (Quast *et al.*, 2013). Hits to chloroplast and mitochondria were observed as expected, but bacteria were virtually absent and similar to controls.

An Illumina NextSeq 500 instrument was used to generate paired 150 bp shotgun reads. Run #1 applied the Illumina High Output kit to libraries BS78 and BS79. Run #1 instrument metrics were: 1.8 pM pool loaded, 1% PhiX spike-in with 1.8% aligned, cluster density 138 K/mm^2, 96% pass filter, and 106 Gbp in 345 M PE reads. Barcode analysis indicated 46% BS78 and 49% BS79. Run #2 applied the Illumina High Output kit to library BS78 only. Run #2 metrics were: 1.8 pM pool loaded, 1% PhiX spike-in with 1% aligned, and 110 Gbp in 360 M PE reads. The resulting FASTQ files contained 101 Gbp in 161 M pairs from BS78 run #1, 169 M pairs from BS79 run #1, and 341 M pairs from BS78 run #2.

Dataset validation

To confirm sugarcane origin of the DNA, the run #1 reads were mapped to available BACs, namely the 608 Kbp of R570 BACs (GenBank accessions KF184657.1 to KF184973.1 (de Setta *et al.*, 2014)). Reads were mapped with bowtie2 (Langmead & Salzberg, 2012) version 2.2.5 with options "-p 4 --no-unal --no-mixed --no-discordant --end-to-end --fast". Both sequencing libraries demonstrated concordant pair mapping rates of 4.1% unique, 27% repeat, and 69% unmapped. Genome coverage analysis was inconclusive; the K-mer frequency distribution computed by Jellyfish (Marçais & Kingsford, 2011) version 2.2.4, with K=17 showed no peak above 1X coverage.

Data availability

The data are available at NCBI SRA under BioProject PRJNA345486, Study SRP091668. Amplified reads from BS78 and BS79 have respective accessions SRR5500242 and SRR5500243. Genomic reads from BS78 have accessions are SRR5500246 and SRR5500247. Genomic reads from BS79 have accession SRR5500249.

Author contributions

Design of experiment: TBS, RS. Sample preparation: KD, DMH. Amplicons: MGT, KJM. Sequencing: KG, KB. Bioinformatics: GS, JM, DMH. Manuscript: JM.

Competing interests

No competing interests were disclosed.

Grant information

This work was funded by US Department of Homeland Security (contract HSHQDC-15-C-B0059).

Acknowledgements

The authors are grateful for assistance from Jack Comstock, Per McCord, and M.D. Islam of USDA-ARS.

References

Asano T, Tsudzuki T, Takahashi S, *et al.*: **Complete nucleotide sequence of the sugarcane (*Saccharum officinarum*) chloroplast genome: a comparative analysis of four monocot chloroplast genomes.** *DNA Res.* 2004; **11**(2): 93–99.
PubMed Abstract | Publisher Full Text

Cardoso-Silva CB, Costa EA, Mancini MC, *et al.*: ***De novo* assembly and transcriptome analysis of contrasting sugarcane varieties.** *PLoS One.* 2014; **9**(2): e88462.
PubMed Abstract | Publisher Full Text | Free Full Text

de Setta N, Monteiro-Vitorello CB, Metcalfe CJ, *et al.*: **Building the sugarcane genome for biotechnology and identifying evolutionary trends.** *BMC Genomics.* 2014; **15**(1): 540.
PubMed Abstract | Publisher Full Text | Free Full Text

Edgar RC: **SEARCH_16S: A new algorithm for identifying 16S ribosomal RNA genes in contigs and chromosomes.** *bioRxiv.* 2017; 124131.
Publisher Full Text

Edmé S, Tai P, Glaz B, *et al.*: **Registration of 'CP 96-1252' sugarcane.** *Crop Sci.* 2005; **45**(1): 423–424.
Publisher Full Text

Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–359.
PubMed Abstract | Publisher Full Text | Free Full Text

Marçais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers.** *Bioinformatics.* 2011; **27**(6): 764–770.
PubMed Abstract | Publisher Full Text | Free Full Text

Okura VK, de Souza RS, de Siqueira Tada SF, *et al.*: **BAC-Pool Sequencing and Assembly of 19 Mb of the Complex Sugarcane Genome.** *Front Plant Sci.* 2016; **7**: 342.
PubMed Abstract | Publisher Full Text | Free Full Text

Quast C, Pruesse E, Yilmaz P, *et al.*: **The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.** *Nucleic Acids Res.* 2013; **41**(Database issue): D590–596.
PubMed Abstract | Publisher Full Text | Free Full Text

Sandhu H, Davidson W: **Sugarcane Cultivars Descriptive Fact Sheet: CP 96-1252, CP 01-1372, and CP 00-1101.** 2016.
Reference Source

Schloss PD, Gevers D, Westcott SL: **Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies.** *PLoS One.* 2011; **6**(12): e27310.
PubMed Abstract | Publisher Full Text | Free Full Text

Shearman JR, Sonthirod C, Naktang C, *et al.*: **The two chromosomes of the mitochondrial genome of a sugarcane cultivar: assembly and recombination analysis using long PacBio reads.** *Sci Rep.* 2016; **6**: 31533.
PubMed Abstract | Publisher Full Text | Free Full Text

Vilela MM, Del Bem LE, Van Sluys MA, *et al.*: **Analysis of Three Sugarcane Homo/Homeologous Regions Suggests Independent Polyploidization Events of *Saccharum officinarum* and *Saccharum spontaneum*.** *Genome Biol Evol.* 2017; **9**(2): 266–278.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 29 June 2017

doi:10.5256/f1000research.12560.r23843

✔ **Jeremy R. Shearman**
National Center for Genetic Engineering and Biotechnology (BIOTEC), Khlong Luang, Thailand

The manuscript describes the generation of whole genome shotgun sequence data from two separate DNA preparation methods. The methods for data generation are clearly described and the sample that was used has ample information about its origins publicly available and referenced. This dataset will be useful for SNP discovery and comparative genomics of sugarcane cultivars.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 30 May 2017

doi:10.5256/f1000research.12560.r22855

✔ **Paulo Arruda**
Center of Molecular Biology and Genetic Engineering (CBMEG), State University of Campinas (UNICAMP), Campinas, Brazil

The data note reported was produced by 150bp paired-end Illumina sequencing of genomic DNA prepared from the sugarcane variety CP-96-1252. A raw data set of 101 Gbps was generated and made public available.The authors did not present assemblage data which would be useful for the research

community interested in sugarcane genomics. Sequence coverage was not stimated but it seems to be under 1X.
Sugarcane commercial varieties are hybrids between Saccharum officinarum and Saccharum spontaneum. These two parents are highly complex polyploids with ploidy varying from 8-12. In general, the hybrids conserve ~75% of the S. officinarum and 15% of S. spontaneum intact. Around 10% of the hybrid genome are chromosomal recombinants between the two species. This complex situation makes it very difficult assembling large non-chimeric contigs especially using short insert shotgun sequencing.

The high quality data set presented in this data note is of value for those interested in recover short gene regions of interest. Because sugarcane genome sequencing dataset is very scarse I recommend the publication of the note presented here as a source of genome data for the sugarcane community.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Discuss this Article

Version 1

Author Response 29 Jun 2017

Jason Miller, JCVI, USA

A draft genome sequence assembly, for a different hybrid of sugarcane, appeared shortly after this paper.

Riaño-Pachón DM and Mattiello L. Draft genome sequencing of the sugarcane hybrid SP80-3280. *F1000Research* 2017, **6**:861 (doi: 10.12688/f1000research.11859.1)

Competing Interests: No competing interests were disclosed.