DATA NOTE

REVISED Draft genome sequencing of the sugarcane hybrid SP80-3280 [version 2; referees: 2 approved]

Diego Mauricio Riaño-Pachón [1,2], Lucia Mattiello[1,3]

[1]Brazilian Bioethanol Science and Technology Laboratory (CTBE), Brazilian Center for Research in Energy and Materials (CNPEM), Campinas, SP, Brazil
[2]Current address: Laboratory of Regulatory Systems Biology, Department of Biochemistry, Institute of Chemistry, University of São Paulo, São Paulo, SP, Brazil
[3]Current address: Functional Genome Laboratory, Department of Genetics, Evolution and Bioagents, Institute of Biology, State University of Campinas, Campinas, SP, Brazil

v2 **First published:** 09 Jun 2017, **6**:861 (doi: 10.12688/f1000research.11859.1)
Latest published: 03 Jul 2017, **6**:861 (doi: 10.12688/f1000research.11859.2)

Abstract
Sugarcane commercial cultivar SP80-3280 has been used as a model for genomic analyses in Brazil. Here we present a draft genome sequence employing Illumina TruSeq Synthetic Long reads. The dataset is available from NCBI BioProject with accession PRJNA272769.

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
REVISED version 2 published 03 Jul 2017		
version 1 published 09 Jun 2017	✓ report	✓ report

1 **Jason Miller**, J. Craig Venter Institute, USA

2 **Chakravarthi Mohan**, Federal University of São Carlos, Brazil

Discuss this article

Comments (0)

Corresponding author: Diego Mauricio Riaño-Pachón (diriano@gmail.com)

Author roles: Riaño-Pachón DM: Conceptualization, Formal Analysis, Methodology, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Mattiello L**: Conceptualization, Methodology, Resources, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Riaño-Pachón DM and Mattiello L. **Draft genome sequencing of the sugarcane hybrid SP80-3280 [version 2; referees: 2 approved]** *F1000Research* 2017, **6**:861 (doi: 10.12688/f1000research.11859.2)

Grant information: This work was supported by institutional funds from CTBE/CNPEM to DMRP and a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant to LM (2012/23345-0). The research was developed with support from CENAPAD-SP (Centro Nacional de Processamento de Alto Desempenho em São Paulo), project UNICAMP/FINEP-MCT.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 09 Jun 2017, **6**:861 (doi: 10.12688/f1000research.11859.1)

REVISED **Amendments from Version 1**

We fixed some spelling mistakes and added information and links about the genome annotation of sugarcane cultivar SP80-3280.

See referee reports

Introduction

Sugarcane is an economically important crop used as source of sugar, ethanol and electricity generation[1]. Sugarcane has a haploid genome of ~1Gpb, however, modern sugarcane cultivars are polyploids derived from interspecific hybridization between *S. officinarum* L. and *S. spontaneum* L., reaching up to 130 chromosomes distributed among ~12 homo(eo)logous groups[2,3], with a total genome size reaching 10Gpb[4]. Its complex genome structure has hampered genome sequencing, assembly and annotation. Partial genomic sequences are available[5–8], as well as transcriptome sequences[9–11], but there are no whole genome assemblies available to date. Here we used the Illumina TruSeq Synthetic Long Read sequencing technology to survey the genome of the polyploid cultivar SP80-3280. The generated long reads, their assembly and genome annotation have been made public and will provide useful information for functional genomics studies.

Materials and methods

The leaf rolls of greenhouse grown, two-month old plants of sugarcane cultivar SP80-3280 (provided by Centro de Tecnologia Canavieira, Piracicaba, São Paulo), were collected and immediately frozen in liquid nitrogen. The plant tissue was ground up to become fine powder, and high molecular weight DNA was extracted from 100 mg of fresh frozen tissue using CTAB (Sigma-Aldrich, USA) and chloroform:isoamyl alcohol (Sigma-Aldrich, USA) as previously described[12]. 6µg of DNA were sent to Illumina (CA, USA) for DNA sequencing using TruSeq Synthetic long read technology[13], through their FastTrack Sequencing Service. Sequencing was performed on an Illumina HiSeq2000 system using paired-end chemistry. Nine long read libraries, each generating approx. 600Mbps, were generated, giving an estimated coverage between 4 and 5 of the monoploid genome. A total of 1,378,917 reads longer than 1.5Kbp, or 5,642,855,018 bases, were generated. The underlying 1,966,604,928 short reads amount to 393,320,985,600bp, which would translate to an estimated coverage of 393x of the haploid genome. The maximum read length was 20,918bp, with 36% of the reads being longer than 4.5Kbp. Possible contaminants were removed by comparison against the NCBI's nucleotide database using BLAST[14], keeping only the long reads with best hits against Viridiplantae, resulting in 1,224,061 useful for assembly. Prior to assembly, long reads originating from mitochondria (NC_008360.1) and chloroplast (NC_005878.2) were excluded using mirabait (http://mira-assembler.sourceforge.net/). Reads longer than 1.5Kbp were assembled using Celera's WGS Assembler v8.2[15], using similar parameters as previously described[13], except for some of the error parameters that were left in their default settings, i.e., '`unitiger=bogart, merSize=31, ovlMinLen=100`', and the parameters `ovlErrorRate, cnsErrorRate, cgwErrorRate, utgGraphErrorRate, utgGraphErrorLimit, utgMergeErrorRate, utgMergeErrorLimit`. A non-redundant assembly was created using CD-HIT[16], merging 100% identical sequences and sub-sequences. RNASeq data previously generated in our group[17] for the same cultivar was exploited for gene prediction using BRAKER1[18] and PASA[19], as well as sugarcane transcript data (ESTs), and *Sorghum bicolor* proteins using Exonerate[20], all gene evidence was integrated to generate a high quality gene prediction set with Evidence Modeller[21], leading to 153,078 predicted protein-coding genes.

Data availability

Raw sequencing data are available at NCBI SRA; the long reads with accession number SRX845504, and the underlying short reads with accessions SRX853961 to SRX853969. The SP80-3280 assembly is available with accession number GCA_002018215.1. All data can be found under the BioProject PRJNA272769. Genome annotation is available from https://figshare.com/projects/Sugarcane_SP80-3280_draft_genome_annotation/22327

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by institutional funds from CTBE/CNPEM to DMRP and a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant to LM (2012/23345-0). The research was developed with support from CENAPAD-SP (Centro Nacional de Processamento de Alto Desempenho em São Paulo), project UNICAMP/FINEP-MCT.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors are grateful to Larissa Prado da Cruz (CTBE/CNPEM) for assistance with molecular biology procedures.

References

1. Long SP, Karp A, Buckeridge MS, *et al.*: **Feedstocks for Biofuels and Bioenergy.** In *Bioenergy & Sustainability: bridging the gaps.* (eds. Souza GM, Victoria RL, Joly CA & Verdade LM), UNESCO. 2015; 302–347.
Reference Source

2. Grivet L, Arruda P: **Sugarcane genomics: depicting the complex genome of an important tropical crop.** *Curr Opin Plant Biol.* 2002; **5**(2): 122–127.
PubMed Abstract | Publisher Full Text

3. D'Hont A: **Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana.** *Cytogenet Genome Res.* 2005; **109**(1–3): 27–33.
PubMed Abstract | Publisher Full Text

4. Le Cunff L, Garsmeur O, Raboin LM, *et al.*: **Diploid/polyploid syntenic shuttle mapping and haplotype-specific chromosome walking toward a rust resistance gene (*Bru1*) in highly polyploid sugarcane (2*n* approximately 12*x* approximately 115).** *Genetics.* 2008; **180**(1): 649–660.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Miller JR, Dilley KA, Harkins DM, *et al.*: **Initial genome sequencing of the sugarcane CP 96-1252 complex hybrid [version 1; referees: 1 approved].** *F1000Res.* 2017; **6**: 688.
Publisher Full Text

6. Grativol C, Regulski M, Bertalan M, *et al.*: **Sugarcane genome sequencing by methylation filtration provides tools for genomic research in the genus *Saccharum*.** *Plant J.* 2014; **79**(1): 162–172.
PubMed Abstract | Publisher Full Text | Free Full Text

7. 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

8. 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

9. Mattiello L, Riaño-Pachón DM, Martins MC, *et al.*: **Physiological and transcriptional analyses of developmental stages along sugarcane leaf.** *BMC Plant Biol.* 2015; **15**: 300.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Hoang NV, Furtado A, Mason PJ, *et al.*: **A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and *de novo* assembly from short read sequencing.** *BMC Genomics.* 2017; **18**(1): 395.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Belesini AA, Carvalho FMS, Telles BR, *et al.*: ***De novo* transcriptome assembly of sugarcane leaves submitted to prolonged water-deficit stress.** *Genet Mol Res.* 2017; **16**(2).
PubMed Abstract | Publisher Full Text

12. Porebski S, Bailey LG, Baum BR: **Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components.** *Plant Mol Biol Rep.* 1997; **15**(1): 8–15.
Publisher Full Text

13. McCoy RC, Taylor RW, Blauwkamp TA, *et al.*: **Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements.** *PLoS One.* 2014; **9**(9): e106689.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Altschul SF, Gish W, Miller W, *et al.*: **Basic local alignment search tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text

15. Myers EW, Sutton GG, Delcher AL, *et al.*: **A Whole-Genome Assembly of *Drosophila*.** *Science.* 2000; **287**(5461): 2196–2204.
PubMed Abstract | Publisher Full Text

16. Fu L, Niu B, Zhu Z, *et al.*: **CD-HIT: accelerated for clustering the next-generation sequencing data.** *Bioinformatics.* 2012; **28**(23): 3150–3152.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Riaño-Pachón DM, Mattiello L, Cruz LP: **Surveying the complex polyploid sugarcane genome sequence using synthetic long reads.** Technical Memorandum Centro Nacional de Pesquisa em Energia e Materiais. 2016.
Publisher Full Text

18. Hoff KJ, Lange S, Lomsadze A, *et al.*: **BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS.** *Bioinformatics.* 2016; **32**(5): 767–9.
PubMed Abstract | Publisher Full Text

19. Haas BJ, Delcher AL, Mount SM, *et al.*: **Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies.** *Nucleic Acids Res.* 2003; **31**(19): 5654–66.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics.* 2005; **6**: 31.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Haas BJ, Salzberg SL, Zhu W, *et al.*: **Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.** *Genome Biol.* 2008; **9**(1): R7.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 21 June 2017

doi:10.5256/f1000research.12814.r23667

✔ **Chakravarthi Mohan**
Department of Genetics and Evolution, Federal University of São Carlos, São Carlos, Brazil

The data note entitled 'Draft genome sequencing of the sugarcane hybrid SP80-3280' is perhaps the first report describing the whole genome of sugarcane, a complex polyploid and its availability in NCBI will be a boon to sugarcane researchers.

The study is well planned, executed and well drafted. The data presented here would be particularly useful for functional genomic studies in sugarcane.

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.

Referee Expertise: Sugarcane genetic engineering, transcriptomics

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.

Author Response 23 Jun 2017

Diego Mauricio Riaño-Pachón, Molecular Biology, University of São Paulo, Brazil

Dear Dr. Mohan,

thanks you for your review of our data note. In version 2 of the note we have added links for the genome annotation in addition to the genome assembly.

Best regards,

Diego

Competing Interests: No competing interests were disclosed.

Referee Report 15 June 2017

doi:10.5256/f1000research.12814.r23398

Jason Miller
J. Craig Venter Institute, Rockville, MD, USA

Summary:

The Data Note, "Draft genome sequencing of the sugarcane hybrid SP80-3280", describes a sugarcane genome assembly that is available at NCBI. The TruSeq method was applied to a monoploid sugarcane cultivar to generate a 1.2 gigabase assembly with a 8433 contig N50 according to GenBank. This is the first sugarcane genome assembly so it will be of interest to the field. This data note is especially useful because it describes the sequence filtering by size, blast, mirabit, and cd-hit prior to release.

Suggestions:

The sentence, “there are not whole genome assemblies available”, probably should say “there are no whole genome assemblies available”. The text could be made clearer by presenting all the statics for underlying short reads before getting to the synthetic long read stats, and by specifying that the blast filter was applied to the long reads. I would appreciate a reference for Celera Assembler, but that is just me.

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.

Referee Expertise: Genome assembly

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.

Author Response 23 Jun 2017

Diego Mauricio Riaño-Pachón, Molecular Biology, University of São Paulo, Brazil

Dear Dr. Miller,

thank you very much for your review of our data note. We have followed your main suggestions, and they are available as version 2 of the data note.

Best regards,

Diego

Competing Interests: No competing interests were disclosed.