Chromosome-scale assembly of the polyploid genome of sugarcane SP80-3280

Gabriely Santos de Oliveira[1]; Diego Mauricio Riaño-Pachón[1 2]

1. Computational, Evolutionary and Systems Biology Laboratory - LabBCES - CENA/USP, Piracicaba, Brazil.
2. National Institute of Science and Technology of Bioethanol - INCT, Butantã, Brazil.

E-mail: gabriely.santos@usp.br; diego.riano@cena.usp.br;

INTRODUCTION

Sugarcane (Saccharum complex) is one of the most important crops worldwide, responsible for over 80% of global sugar production and widely used in bioethanol and bioenergy industries. However, its genome is exceptionally challenging to assemble due to both its enormous size >10 gigabases (Gb) and its highly complex polyploid nature.

Modern sugarcane cultivars, such as SP80-3280, are the result of interspecific hybridization between Saccharum officinarum and S. spontaneum, followed by backcrossing and selection. As a result, sugarcane is a highly heterozygous, aneuploid, and autoallopolyploid species, carrying multiple copies of each chromosome from both ancestral lineages

This complexity leads to severe challenges in genome assembly: Extensive sequence redundancy across homeologous chromosomes, high allelic variation and haplotype diversity and the difficulty in resolving structural variations and phase-separated haplotypes.

Currently, two chromosome-scale assemblies are being scaffolded: one with 104 pseudochromosomes and another with 112, both guided by cytogenetic evidence and structural collinearity.

METHODS

SEQUENCING

HiFi reads (312.4 Gb from PacBio Sequel II) and **Hi-C data** (~450 Gb Illumina PE150 from 4 RE libraries) were generated for the SP80-3280 sugarcane genome.

ASSEMBLY

HiFiASM
phased contig-level assembly
HiFi + Hi-C reads.

→ **Assembly Evaluation**
Quast and Compleasm.

SCAFFOLDING

HapHiC
Hi-C data
pseudochromosomes.

→ Scaffold **accuracy** is currently being reviewed and manually curated.

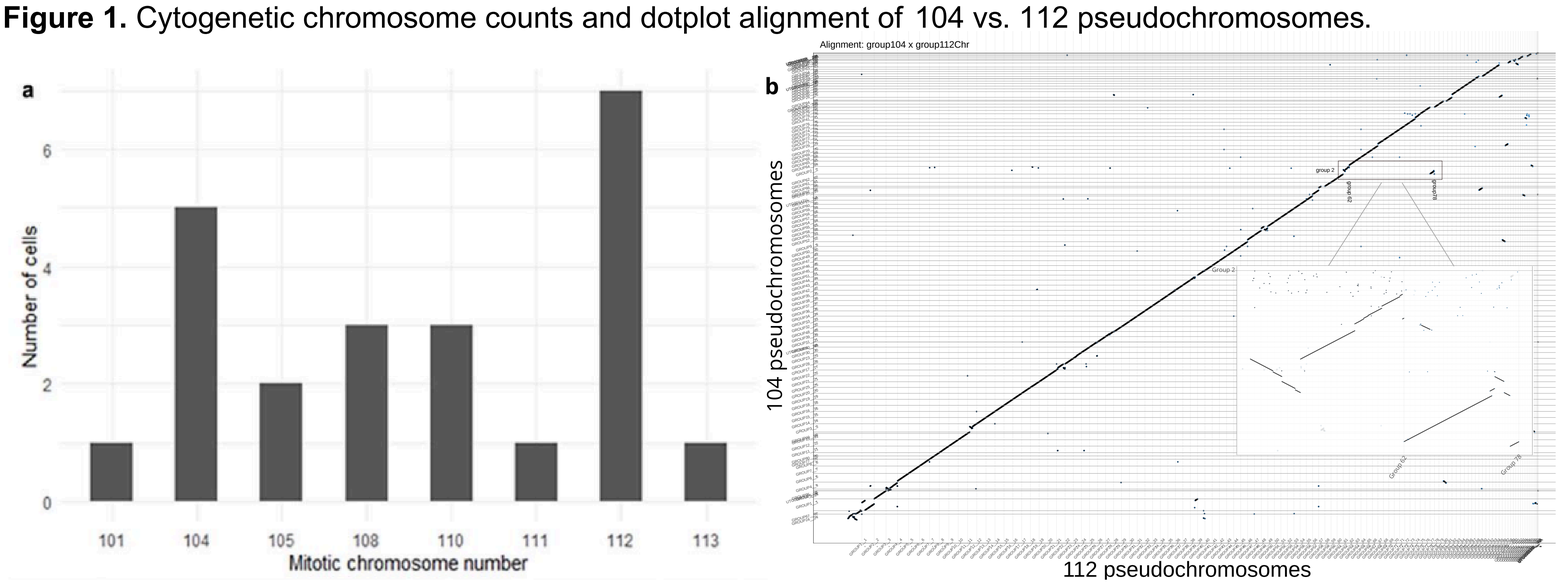
GENOME ANNOTATION

The new assembly is being **compared** to previous versions (2017 and 2019) and ancestral species (S. *officinarum* and S. *spontaneum*).

The genome is currently being **annotated** with a focus on key regulatory genes such as LFY, PHYC, and TOR.

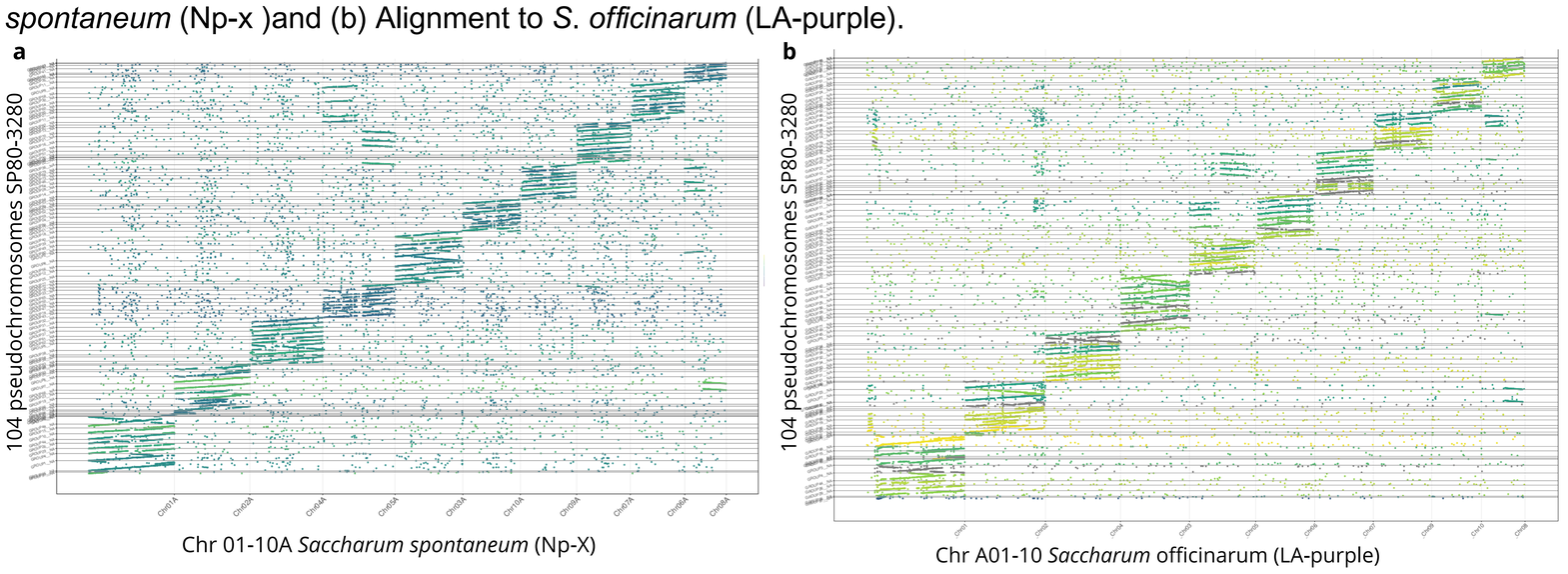
RESULTS

Figure 1. Cytogenetic chromosome counts and dotplot alignment of 104 vs. 112 pseudochromosomes.



a. Mitotic counts support a modal chromosome number of 112 in SP80-3280 (Oliveira et al., 2022); b. Dotplot comparison between Assembly 104 (Asm104) and Assembly 112 (Asm112) shows high collinearity, validating the chromosome-scale structure. Additionally, segmental alignments from group 2 (Asm104) to groups 62 and 78 (from Asm112) suggest a possible chromosome split between assemblies. When compared against the *S. spontaneum* and *S. officinarum* assemblies, the sequence represented by group 2 (Asm104) appears to be the correct one. Further analysis is ongoing to characterize this region in more detail.

Figure 2. Whole-genome alignment of SP80-3280 assembly-104 to ancestral genomes:(a) Alignment to *S. spontaneum* (Np-x)and (b) Alignment to *S. officinarum* (LA-purple).



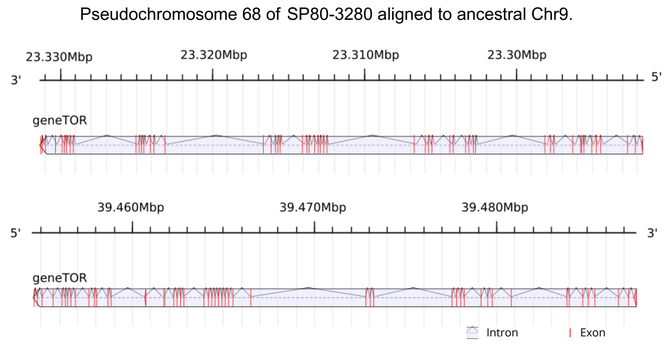
Whole-genome alignment of SP80-3280 Asm104 to ancestral genomes. (a) Alignment to *S. spontaneum;* (b) Alignment to *S. officinarum.* Dotploty alignments were performed individually against each ancestral chromosome, excluding duplicated haplotypes, to highlight syntenic structure and collinearity.

Table 1. Assembly quality and completeness of sugarcane genomes (2017, 2019, and 2024).

QUAST	CTBE_SP803280 **(2017)**	ASM869266v1 **(2019)**	**SP80-3280 Chr104 (2024)**
Total Contigs	199.028	398.353	35.302
Contigs ≥10 Mbp	0	0	**102**
Largest contig	115.91 Kbp	468.01 Kbp	171.94 Mbp
Total length	1.17 Gbp	4.01 Gbp	**9.12 Gbp**
GC (%)	43,04%	43,71%	44,73%
N50	8.45 Kbp	14.00 Kbp	**70.43 Mbp**
N75	5.20 Kbp	8.68 Kbp	33.96 Mbp
Compleasm C	C:70,15%, 2270	C;86,65%, 2804	**C:99,69%, 3226**
F	F:17.24%, 558	F:8.56%, 277	F:0.22%, 7
M	M:12.61%, 408	M:4.76%, 154	M:0.09%, 3

Metrics from QUAST v5.0.2 and Compleasm v0.2.6 using the *Liliopsida* BUSCO lineage set.

Figure 3. Mapping of TOR1 gene copies across pseudochromosomes in SP80-3280.



Mapping of TOR gene sequences against the assembly 104 (group 68) SP80-3280 revealed two distinct loci. Although preliminary, this may suggest the presence of more than one TOR copy in this genome. Further validation is ongoing. Previous studies have described TOR as a single-copy gene in sugarcane (Vilela et al., 2017).