Genome report: A duplication lost in sugarcane hybrids revealed by chloroplast genome assembly of wild species *Saccharum officinarum*

Deise Paes[*], Filipe Pereira Matteoli[*], Thiago Motta Venancio[*], Paulo Cavalcanti Gomes Ferreira[†], Clicia Grativol[*]

[*] Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Av. Alberto Lamego, 2000, P5-228, Parque Califórnia, Campos dos Goytacazes-RJ, 28013-602, Brazil.

† Laboratório de Biologia Molecular de Plantas, Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Cidade Universitária, Avenida Carlos Chagas Filho, 373, CCS, Bl.L-29ss, Rio de Janeiro – RJ, 21941-599, Brazil.

Running title: CpDNA assembly of *S. officinarum*

Keywords: chloroplast genome; sugarcane; wild species; WGS; *Saccharum officinarum*

Corresponding author:

Clicia Grativol

Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Av. Alberto Lamego, 2000, P5-228, Parque Califórnia, Campos dos Goytacazes-RJ, 28013-602, Brazil

cgrativol@uenf.br, +55222749-7107

ABSTRACT

Sugarcane is a crop of paramount importance for sustainable energy. Modern sugarcane cultivars are derived from interspecific crosses between the two wild species *Saccharum officinarum* and *Saccharum spontaneum* and this event occurred very early in the sugarcane domestication history. This hybridization allowed the generation of cultivars with complex aneuploidy genomes containing 100–130 chromosomes that are unequally inherited - ~80% from *S. officinarum*, ~10% from *S. spontaneum* and ~10% from inter-specific crosses. Several studies have highlighted the importance of chloroplast genomes (cpDNA) to investigate hybridization events in plant lineages. Few sugarcane cpDNAs have been assembled and published, including those from sugarcane hybrids. However, cpDNAs of wild Saccharum species remains unexplored. In the present study, we used whole-genome sequencing data to survey the chloroplast genome of the wild sugarcane species *S. officinarum*. Illumina sequencing technology was used for assembly 142,234 bp of *S.officinarum* cpDNA with 2,065,893 reads and 1043x of coverage. The analysis of the *S. officinarum* cpDNA revealed a notable difference in the LSC region of wild and cultivated sugarcanes. Chloroplasts of sugarcane cultivars showed a loss of a duplicated fragment with 1,031 bp in the beginning of the LSC region, which decreased the chloroplast gene content in hybrids. Based on these results, we propose the comparative analysis of organelle genomes as a very important tool for deciphering and understanding hybrid Saccharum lineages.

INTRODUCTION

Sequencing of organelle genomes is an important tool in molecular and evolutionary studies (Wolf *et al.* 2011). In addition to the nuclear genome, plants have mitochondrial (mtDNA) and chloroplast genomes (cpDNA), which can allow a broad

analysis on specific species (Xu *et al.* 2015). The size of cpDNAs of land plants ranges from 100 to 160 kb, with around 100 to 120 highly conserved genes (Wicke *et al.* 2011; Olejniczak *et al.* 2016). The features of the cpDNAs are also helpful in phylogenetic studies and to develop genetic markers (Bock and Khan 2004; Jansen *et al.* 2007; Ravi *et al.* 2008; Wu and Ge 2012). Further, several studies highlighted the importance of cpDNAs to investigate hybridization events than nuclear genomes; cpDNAs allow the analysis of organelle sharing patterns between species due to their slow rate of evolution, non-recombinant nature, easy haplotype detection and predominantly uniparental inheritance (Wu *et al.* 2010; Smith 2015; Zhu *et al.* 2016; Szczecińska *et al.* 2017; Xiao-Ming *et al.* 2017). Many studies have been conducted with chloroplast genomes to identify the history of plant lineages (Marí-ordóñez *et al.* 2013; Rousseau-Gueutin *et al.* 2015; Cho *et al.* 2016; Shetty *et al.* 2016; Yang *et al.* 2016; Asaf *et al.* 2017). As an example, the sequencing of chloroplast genomes of *Solanum commersonii* and *Solanum tuberosum* revealed indel markers that can distinguish chlorotypes and maternal inheritance of these organelles in hybrids (Cho *et al.* 2016).

Many species from the *Saccharum* genus (Poaceae) have been widely used in sugar production due to their remarkable sucrose storage capacity. Due to its tropical and subtropical distribution, sugarcane has probably been first established at New Guinea and Indonesia (Grivet *et al.* 2006). *S. officinarum* has a chromosome number of 2n = 80 and is known as "noble" sugarcane, mainly due to its high sucrose content, large and thick low-fiber stalks (Cheavegatti-Gianotto et al., 2011). Despite these key agronomic traits, this species is water-intensive, susceptible to diseases and requires high soil fertility. In the end of 19 century, a cross between *Saccharum spontaneum* and *Saccharum officinarum* resulted in a hybrid that was then backcrossed with *Saccharum officinarum*. The introgression of a small part of the *S. spontaneum* genome into a

predominantly *S. officinarum* genome resulted in modern hybrids (*Saccharum spp*.) with better yields, high sucrose content and ability to cope some biotic and abiotic stresses. These hybrids were critical for the development of the sugar trade (Grivet and Arruda 2002; Moore 2005; Cheavegatti-Gianotto *et al.* 2011). Modern sugarcane cultivars have complex and aneuploidy nuclear genomes. Few sugarcane cpDNAs have been assembled and published, including those from the sugarcane hybrids *Saccharum* spp. Q155 (Hoang *et al.* 2015), *Saccharum* spp. NCo 310 (Asano *et al.* 2004), *Saccharum* spp. SP80-3280 (Calsa Júnior *et al.* 2004) and *Saccharum* spp. RB867515 (Vidigal *et al.* 2016). However, cpDNAs of wild Saccharum species remains unexplored. In the present study, we used whole-genome sequencing data to survey the chloroplast genome of the wild sugarcane species *S. officinarum*.

METHODS AND MATERIALS

Plant material

Young leaves from *S. officinarum* accession 82-72 maintained in the germplasm collection of Instituto Agronômico de Campinas (Ribeirão Preto, Brazil) were used for DNA analysis. According to Kuijper's leaf numbering system for sugarcane (Cheavegatti-Gianotto *et al.* 2011), leaf -2 tissue was used to subsequent DNA extraction and sequencing.

DNA extraction and sequencing

Total genomic DNA was extracted from leaves using the CTAB method (Doyle and Doyle 1987) with minor modifications. The quality of DNA was estimated using

Thermo Scientific *NanoDrop*™ 2000c Spectrophotometer. Total DNA (~20ug) was sequenced on the Illumina GAII machine using the paired-end 100 cycle protocol.

De novo assembly of chloroplast using genomic DNA reads

The sequencing reads were initially filtered to retain those with 90% of bases having quality scores greater than or equal to 20 (Q20) using FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). After quality filtering, we performed a BLASTN (Altschul *et al.* 1997) search (e-value $\leq 10^{-4}$) with chloroplast sequences from *Saccharum* hybrid cultivar NCo 310 (NC_006084.1), *Saccharum* hybrid cultivar SP-80-3280 (AE009947.2), *Sorghum bicolor* (NC_008602.1), *Zea mays* (NC_001666.2), *Miscanthus sinensis* (NC_028721.1), *Oryza sativa* (KT289404.1) and *Setaria italica* (KJ001642.1). The reads aligned to cpDNAs were tested on VelvetOptimiser (https://github.com/tseemann/VelvetOptimiser) with k-mer range from 29 to 87. Genome assembly was performed with SPADES (Bankevich *et al.* 2012) using the following parameters: 53, 69 and 77 of k-mers; 70 of coverage cutoff and careful parameter. The SSPACE (Boetzer *et al.* 2011) was run with default parameters on SPADES assembled contigs. The assembled chloroplast was compared with cpDNA from Saccharum hybrid cultivar Q155 using BLASTN, annotated with GeSeq (Tillich *et al.* 2017) and the resulting Genbank file was visualized on OrganellarGenomeDRAW (Lohse *et al.* 2013).

Data availability

The sequence data from whole genome shotgun of sugarcane wild species *S. officinarum* have been submitted to the NCBI Sequence Read Archive under accession SRX313496. The assembled chloroplast genome sequence is available at NCBI

Genbank with accession number MF140336 (http://www.ncbi.nlm.nih.gov/nuccore/MF140336).

RESULTS AND DISCUSSION

A total 297,637,906 whole-genome shotgun reads of *S. officinarum* were sequenced using an Illumina GAII platform. Quality filtered reads were screened for similarity with known chloroplast sequences (see methods for details), which resulted in 2,065,893 reads that were used to assemble the 142,234 bp *S. officinarum* cpDNA at 1043x coverage (Table 1). Five scaffolds were assembled, the largest one with 106,869 bp (Table 1). This genome has 1,052 bp more than the hybrid cultivar *Saccharum spp.* SP80-3280, reported to have 141,182 bp (Calsa Júnior *et al.* 2004). The *S. officinarum* cpDNA has four main regions: LSC and SSC, with 84,080 bp and 12,576 bp, respectively and; the inverted repeats, IRa and IRb, with 22,789 bp each (Figure 1). Seventy-two genes were annotated, out of which 25 are protein-coding genes, 40 tRNA genes, four rRNA and three other genes (cssa, cemA and infA). In the inverted region, there are 20 duplicate genes: ten tRNA, four rRNA and six protein-coding genes. The IR junction with LSC is between the *rpl22* and *trnH-rps19* gene cluster. Accordingly, the *trnH-rps19* gene cluster is present close to in the IR/LSC junction region in other monocotyledons species chloroplasts (Wang *et al.* 2008).

The analysis of of the *S. officinarum* cpDNA revealed a notable difference in the LSC region of wild and cultivated sugarcanes. Chloroplasts of sugarcane cultivars such as *Saccharum* spp. Q155 (Hoang *et al.* 2015), *Saccharum* spp. NCo 310 (Asano *et al.* 2004), *Saccharum* spp. SP80-3280 (Calsa Júnior *et al.* 2004) and *Saccharum* spp. RB867515 (Vidigal *et al.* 2016) showed a loss of a duplicated fragment with 1,031 bp in the beginning of the LSC region. In comparison with those cultivars' chloroplasts, *S.*

officinarum has an insertion of 10 bp inside the *rpl23-F* gene and two copies of *orf137*, *trnT*, *orf74* and *rps19* genes. Like the NCo310 chloroplast, *S. officinarum* chloroplast has an intron in the middle of the *rpl2* gene. Based on these results, we propose the comparative analysis of organelle genomes as a very important tool for deciphering and understanding hybrid Saccharum lineages.

ACKNOWLEDGEMENTS

We are grateful to Instituto Agronômico de Campinas for providing the wild species plant materials. We are thankful to FAPERJ (Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for the financial support.

REFERENCES

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Asaf, S., M. Waqas, A. L. Khan, M. A. Khan, S. Kang *et al.*, 2017 The Complete Chloroplast Genome of Wild Rice (Oryza minuta) and Its Comparison to Related Species. Front. Plant Sci. 8: 1–15.

Asano, T., T. Tsudzuki, S. Takahashi, H. Shimada, and K. Kadowaki, 2004 Complete Nucleotide Sequence of the Sugarcane (Saccharum. 99: 93–99.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin *et al.*, 2012 SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell

Sequencing. J. Comput. Biol. 19: 455–477.

Bock, R., and M. S. Khan, 2004 Taming plastids for a green future. Trends Biotechnol. 22: 311–318.

Boetzer, M., C. V Henkel, H. J. Jansen, D. Butler, and W. Pirovano, 2011 Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27: 578–9.

Calsa Júnior, T., D. M. Carraro, M. R. Benatti, A. C. Barbosa, J. P. Kitajima *et al.*, 2004 Structural features and transcript-editing analysis of sugarcane (Saccharum officinarum L.) chloroplast genome. Curr. Genet. 46: 366–373.

Cheavegatti-Gianotto, A., H. M. C. de Abreu, P. Arruda, J. C. Bespalhok Filho, W. L. Burnquist *et al.*, 2011 Sugarcane (Saccharum X officinarum): A Reference Study for the Regulation of Genetically Modified Cultivars in Brazil. Trop. Plant Biol. 4: 62–89.

Cho, K. S., K. S. Cheon, S. Y. Hong, J. H. Cho, J. S. Im *et al.*, 2016 Complete chloroplast genome sequences of Solanum commersonii and its application to chloroplast genotype in somatic hybrids with Solanum tuberosum. Plant Cell Rep. 1–11.

Daniell, H., C.-S. Lin, M. Yu, and W.-J. Chang, 2016 Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 17: 134.

Daniels, J., and B. T. Roach, 1987 Taxonomy and evolution., pp. 7–84 in *Sugarcane Improvement Through Breeding*, edited by D. J. Heinz. Elsevier, Amsterdam, The Netherlands.

Doyle, J. J. ., and J. L. Doyle, 1987 A rapid DNA isolation method for small quantities of fresh tissues. Phytochem. Bull. 19:.

Grivet, L., and P. Arruda, 2002 Sugarcane genomics: depicting the complex genome of an important tropical crop. Curr. Opin. Plant Biol. 5: 122–7.

Grivet, L., J. Glaszmann, and A. D'Hont, 2006 Molecular evidence of sugarcane evolution and domestication. Darwin's Harvest. New Approaches to Orig. Evol. Conserv. Crop. 49–66.

Hoang, N. V., A. Furtado, R. B. McQualter, and R. J. Henry, 2015 Next generation sequencing of total DNA from sugarcane provides no evidence for chloroplast heteroplasmy. New Negatives Plant Sci. 1–2: 33–45.

Jansen, R. K., Z. Cai, L. a Raubeson, H. Daniell, C. W. Depamphilis *et al.*, 2007 Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl. Acad. Sci. U. S. A. 104: 19369–19374.

Lohse, M., O. Drechsel, S. Kahlau, and R. Bock, 2013 OrganellarGenomeDRAW--a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. 41: 575–581.

Marí-ordóñez, A., A. Marchais, M. Etcheverry, A. Martin, V. Colot *et al.*, 2013 Reconstructing de novo silencing of an active plant retrotransposon.

Moore, P. H., 2005 Integration of sucrose accumulation processes across hierarchical scales: towards developing an understanding of the gene-to-crop continuum. F. Crop. Res. 92: 119–135.

Olejniczak, S. A., E. ??ojewska, T. Kowalczyk, and T. Sakowicz, 2016 Chloroplasts: state of research and practical applications of plastome sequencing. Planta 244: 517–527.

Ravi, V., J. P. Khurana, A. K. Tyagi, and P. Khurana, 2008 An update on chloroplast genomes. Plant Syst. Evol. 271: 101–122.

Rousseau-Gueutin, M., S. Bellot, G. E. Martin, J. Boutte, H. Chelaifa *et al.*, 2015 The chloroplast genome of the hexaploid Spartina maritima (Poaceae, Chloridoideae):

Comparative analyses and molecular dating. Mol. Phylogenet. Evol. 93: 5–16.

Shetty, S. M., M. U. Shah, K. Makale, Y. Mohd-yusuf, N. Khalid *et al.*, 2016 Complete Chloroplast Genome Sequence of Musa balbisiana Corroborates Structural Heterogeneity of Inverted Repeats in Wild Progenitors of Cultivated Bananas and Plantains. Plant Genome 9: 1–14.

Smith, D. R., 2015 Mutation rates in plastid genomes: They are lower than you might think. Genome Biol. Evol. 7: 1227–1234.

Szczecińska, M., G. Łazarski, K. Bilska, and J. Sawicki, 2017 The complete plastid genome and nuclear genome markers provide molecular evidence for the hybrid origin of Pulsatilla × hackelii Pohl . Turk. J. Botany 41: 1–9.

Tillich, M., P. Lehwark, T. Pellizzer, E. S. Ulbricht-Jones, A. Fischer *et al.*, 2017 GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 22: 97–104.

Vidigal, P. M. P., A. S. G. Coelho, E. Novaes, M. H. P. Barbosa, L. A. Peternelli *et al.*, 2016 Complete Chloroplast Genome Sequence and Annotation of the Saccharum hybrid cultivar RB867515. Genome Announc 4: e01157-16.

Wang, R.-J., C.-L. Cheng, C.-C. Chang, C.-L. Wu, T.-M. Su *et al.*, 2008 Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. BMC Evol. Biol. 8: 36.

Wicke, S., G. M. Schneeweiss, C. W. dePamphilis, K. F. Müller, and D. Quandt, 2011 The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. Plant Mol. Biol. 76: 273–297.

Wolf, P. G., J. P. Der, A. M. Duffy, J. B. Davidson, A. L. Grusz *et al.*, 2011 The evolution of chloroplast genes and genomes in ferns. Plant Mol. Biol. 76: 251–261.

Wu, F.-H., M.-T. Chan, D.-C. Liao, C.-T. Hsu, Y.-W. Lee *et al.*, 2010 Complete

chloroplast genome of Oncidium Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. BMC Plant Biol. 10: 68.

Wu, Z. Q., and S. Ge, 2012 The phylogeny of the BEP clade in grasses revisited: Evidence from the whole-genome sequences of chloroplasts. Mol. Phylogenet. Evol. 62: 573–578.

Xiao-Ming, Z., W. Junrui, F. Li, L. Sha, P. Hongbo *et al.*, 2017 Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants. Sci. Rep. 7: 1555.

Xu, J., Q. Liu, W. Hu, T. Wang, Q. Xue *et al.*, 2015 Dynamics of chloroplast genomes in green plants. Genomics.

Yang, Y., T. Zhou, D. Duan, J. Yang, L. Feng *et al.*, 2016 Comparative Analysis of the Complete Chloroplast Genomes of Five Quercus Species. Front. Plant Sci. 7: 959.

Zhu, A., W. Guo, S. Gupta, W. Fan, and J. P. Mower, 2016 Evolutionary dynamics of the plastid inverted repeat: The effects of expansion, contraction, and loss on substitution rates. New Phytol. 209: 1747–1756.

Table 1 - Summary of *S. officinarum* chloroplast genome assembly statistics.

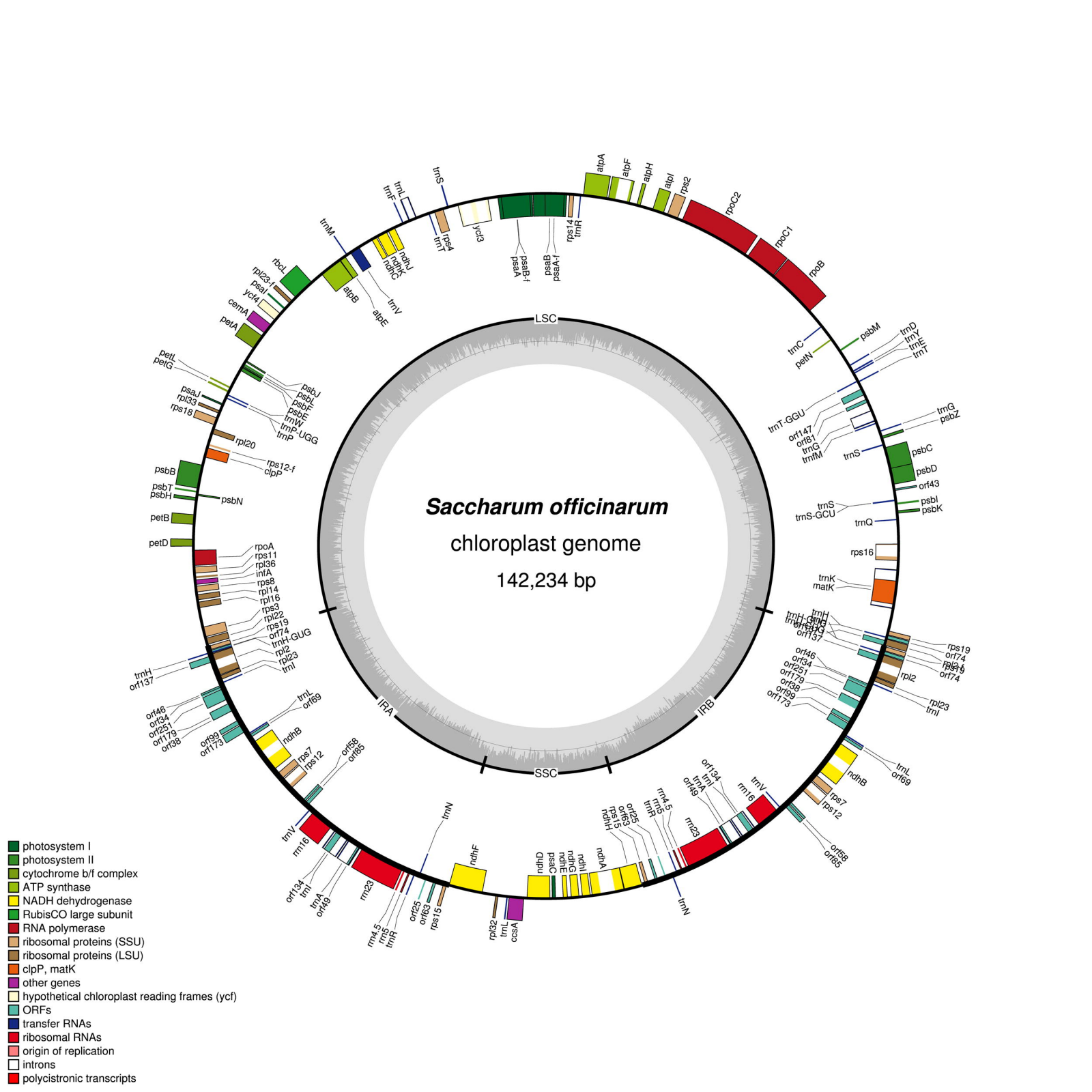
Assembly parameters	Value
Quality filtered reads	151,392,928
Number of reads*	2,065,893
Total length	142,234
Number of scaffolds	5
Largest scaffold (bp)	106,869
% GC	37.41
N50	106,869
Estimated coverage**	1043x

*Assembly input

**Coverage based on Q155 chloroplast size

FIGURE LEGEND

Figure 1- Graphic representation of *S. officinarum* chloroplast genome. Genes and tRNAs elements are identified as coloured boxes. The genes transcription direction is indicated by gray arrows. The locations of large and small single-copy regions and, the pair of inverted repeats (IRa and IRb) are shown in the inner circle. The darker gray color in the inner circle corresponds to the GC content, and the lighter gray color corresponds to the AT content.


Saccharum officinarum
chloroplast genome
142,234 bp
LSC
IRA
IRB
SSC
photosystem I
photosystem II
cytochrome b/f complex
ATP synthase
NADH dehydrogenase
RubisCO large subunit
RNA polymerase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
other genes
hypothetical chloroplast reading frames (ycf)
ORFs
transfer RNAs
ribosomal RNAs
origin of replication
introns
polycistronic transcripts