

The sugarcane mitochondrial genome: assembly, phylogenetics and transcriptomics (#35349)

1

First submission

Guidance from your Editor

Please submit by **8 Apr 2019** for the benefit of the authors (and your \$200 publishing discount).



Structure and Criteria

Please read the 'Structure and Criteria' page for general guidance.



Custom checks

Make sure you include the custom checks shown below, in your review.



Raw data check

Review the raw data. Download from the location [described by the author](#).



Image check

Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files

Download and review all files from the [materials page](#).

10 Figure file(s)

6 Table file(s)

1 Raw data file(s)

! Custom checks

Patent checks



Have you checked the [competing interests statement](#)?



Are there any undeclared competing interests? ([patent policy](#))

DNA data checks



Have you checked the authors [data deposition statement](#)?



Can you access the deposited data?



Has the data been deposited correctly?



Is the deposition information noted in the manuscript?



Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. BASIC REPORTING
2. EXPERIMENTAL DESIGN
3. VALIDITY OF THE FINDINGS
4. General comments
5. Confidential notes to the editor

 You can also annotate this PDF and upload it as part of your review

When ready [submit online](#).

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your [guidance page](#).





BASIC REPORTING

-  Clear, unambiguous, professional English language used throughout.
-  Intro & background to show context. Literature well referenced & relevant.
-  Structure conforms to [PeerJ standards](#), discipline norm, or improved for clarity.
-  Figures are relevant, high quality, well labelled & described.
-  Raw data supplied (see [PeerJ policy](#)).

EXPERIMENTAL DESIGN

-  Original primary research within [Scope of the journal](#).
-  Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
-  Rigorous investigation performed to a high technical & ethical standard.
-  Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

-  Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
-  Data is robust, statistically sound, & controlled.
-  Speculation is welcome, but should be identified as such.
-  Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

3



The best reviewers use these techniques

Tip

Support criticisms with evidence from the text or from other sources

Example

Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.

Give specific suggestions on how to improve the manuscript

Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).

Comment on language and grammar issues

The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult.

Organize by importance of the issues, and number your points

1. Your most important issue
2. The next most important item
3. ...
4. The least important points

Please provide constructive criticism, and avoid personal opinions

I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC

Comment on strengths (as well as weaknesses) of the manuscript

I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.

The sugarcane mitochondrial genome: assembly, phylogenetics and transcriptomics

Dyfed Lloyd Evans^{Corresp., 1, 2, 3}, Thandekile Hlongwane¹, Shailesh V Joshi^{1, 4}, Diego M Riaño Pachón⁵

¹ Plant Breeding, South African Sugarcane Research Institute, Durban, KwaZulu Natal, South Africa

² Cambridge Sequence Services (CSS), Cambridge, Cambridgeshire, United Kingdom

³ Department of Computer Sciences, University Cheikh Anta Diop of Dakar, Dakar, Dakar, Senegal

⁴ College of Agriculture Engineering and Science, University of Kwa-Zulu Natal, Durban, KwaZulu Natal, South Africa

⁵ Computational, Evolutionary and Systems Biology Laboratory, Center for Nuclear Energy in Agriculture, University of Sao Paulo, Piracicaba, Sao Paulo, Brazil

Corresponding Author: Dyfed Lloyd Evans

Email address: dyfed.sa@gmail.com

Background. Chloroplast genomes provide insufficient phylogenetic information to distinguish between closely-related sugarcane cultivars. The mitochondrial genome of plants is much larger and more plastic and could contain increased phylogenetic signals. We attempted to assemble a reference mitochondrion with Illumina TruSeq synthetic long reads. Based on this assembly we also analyzed the mitochondrial transcriptomes of sugarcane and sorghum and to improve the annotation of the sugarcane mitochondrion.

Methods. Mitochondrial genomes were assembled from genomic read pools using a bait and assemble methodology. The mitogenome was exhaustively annotated using blast and transcript datasets were mapped with HISAT2 prior to analysis with the Integrated Genome Viewer.

Results. The sugarcane mitochondrion is comprised of independent chromosomes, which cannot recombine. Based on the reference assembly from the sugarcane cultivar SP80-3280 the mitogenomes of three additional cultivars were assembled (with the SP70-1143 assembly utilizing both genomic and transcriptomic data). We demonstrate that the sugarcane plastome is completely transcribed and we assembled the chloroplast of SP80-3280 using transcriptomic data only. Phylogenomic analysis using mitogenomes allow closely related sugarcane cultivars to be distinguished and supports the discrimination between *Saccharum officinarum* and *Saccharum cultum* as modern sugarcane's female parent. From whole chloroplast comparisons, we demonstrate that modern sugarcane arose from a limited number of *S. cultum* female founders. Transcriptomic and spliceosomal analyses reveal that the two chromosomes of the sugarcane mitochondrion are combined at the transcript level and that splice sites occur more frequently within gene coding regions than without. We reveal a potential cytoplasmic male sterility factor in the sugarcane mitochondrion.

Conclusion. Transcript processing in the sugarcane mitochondrion is highly complex with diverse splice events, the majority of which span the two chromosomes. PolyA baited transcripts are consistent with the use of polyadenylation for transcript degradation. For the first time we annotate a potential cytoplasmic male sterility factor within the sugarcane mitochondrion and demonstrate that sugarcane possesses all the molecular machinery required for cytoplasmic male sterility and rescue. We also demonstrate that mitogenomes can be used to perform phylogenomic studies on sugarcane cultivars.

The Sugarcane Mitochondrial Genome: Assembly, Phylogenetics and Transcriptomics.

Dyfed Lloyd Evans^{1,2,3}, Thandekile Hlongwane¹, Shailesh Vinay Joshi^{1,4} and Diego M. Riaño-Pachón⁵

¹ South African Sugarcane Research Institute, 170 Flanders Drive, Private Bag X02, Mount Edgecombe, Durban, 4300, South Africa

² Department of Computer Science, Université Cheikh Anta Diop de Dakar, BP 5005, Dakar, Sénégal

³ Cambridge Sequence Services (CSS), Waterbeach, Cambridge, CB25 9TL, UK

⁴ School of Life Sciences, College of Agriculture, Engineering and Science, University of Kwa-Zulu Natal, Private Bag X54001, Durban, 4000, South Africa

⁵ Computational, Evolutionary and Systems Biology Laboratory, Center for Nuclear Energy in Agriculture, University of São Paulo, Piracicaba, SP, Brazil

Corresponding Author:

Dyfed Lloyd Evans

Email address: dyfed.sa@gmail.com

20 Abstract

21 **Background.** Chloroplast genomes provide insufficient phylogenetic information to
 22 distinguish between closely-related sugarcane cultivars. The mitochondrial genome of
 23 plants is much larger and more plastic and could contain increased phylogenetic signals.
 24 We attempted to assemble a reference mitochondrion with Illumina TruSeq synthetic long
 25 reads. Based on this assembly we also analyzed the mitochondrial transcriptomes of
 26 sugarcane and sorghum and to improve the annotation of the sugarcane mitochondrion.

27 **Methods.** Mitochondrial genomes were assembled from genomic read pools using a bait
 28 and assemble methodology. The mitogenome was exhaustively annotated using blast and
 29 transcript datasets were mapped with HISAT2 prior to analysis with the Integrated
 30 Genome Viewer.

31 **Results.** The sugarcane mitochondrion is comprised of independent chromosomes, which
 32 cannot recombine. Based on the reference assembly from the sugarcane cultivar SP80-
 33 3280 the mitogenomes of three additional cultivars were assembled (with the SP70-1143
 34 assembly utilizing both genomic and transcriptomic data). We demonstrate that the sugarcane
 35 plastome is completely transcribed and we assembled the chloroplast of SP80-3280 using
 36 transcriptomic data only. Phylogenomic analysis using mitogenomes allow closely related
 37 sugarcane cultivars to be distinguished and supports the discrimination between *Saccharum*
 38 *officinatum* and *Saccharum cultum* as modern sugarcane's female parent. From whole
 39 chloroplast comparisons, we demonstrate that modern sugarcane arose from a limited number of
 40 *S. cultum* female founders. Transcriptomic and spliceosomal analyses reveal that the two
 41 chromosomes of the sugarcane mitochondrion are combined at the transcript level and that splice

sites occur more frequently within gene coding regions than without. We reveal a potential cytoplasmic male sterility factor in the sugarcane mitochondrion.

Conclusion. Transcript processing in the sugarcane mitochondrion is highly complex with diverse splice events, the majority of which span the two chromosomes. PolyA baited transcripts are consistent with the use of polyadenylation for transcript degradation. For the first time we annotate a potential cytoplasmic male sterility factor within the sugarcane mitochondrion and demonstrate that sugarcane possesses all the molecular machinery required for cytoplasmic male sterility and rescue. We also demonstrate that mitogenomes can be used to perform phylogenomic studies on sugarcane cultivars.

Keywords: mitochondria; plastomes; sugarcane; phylogenetics; sugarcane origins; *Saccharum cultum*; cytoplasmic male sterility

Introduction

Sugarcane ranks amongst the top-ten crop species worldwide. Sugarcane also provides between 60 and 70% of total world sugar output and is a major source of bioethanol (Reddy et al., 2008). *Saccharum officinarum* L. is the type species for genus *Saccharum* L. Genus *Saccharum*, in the broad sense, (*sensu lato*) consists of up to 36 species according to Kew's GrassBase (Clayton et al., 2006) or 22 validated species according to Tropicos (<http://tropicos.org/Home.aspx>). However, recent findings indicate that many of these species belong in different genera (Lloyd Evans, Joshi & Wang, 2019) and that *Saccharum sensu stricto* (*s.s.*) [in the strict sense], consists of only four true species: *Saccharum spontaneum* L., *Saccharum robustum* Brandes & Jeswiet ex Grassl, *Saccharum officinarum* and *Saccharum cultum* (Lloyd Evans & Joshi, 2016).

Saccharum officinarum has a centre of diversity in New Guinea (Daniels & Roach, 1987), whilst *Saccharum spontaneum* is distributed from North Africa through to New Guinea, with a centre of diversity in India (Sobhakumari, 2013). Before the 1780s, all sugarcanes arose from essentially sterile wild hybrids of *Saccharum officinarum* and *Saccharum spontaneum* (Artschwager & Brandes, 1958; Irvine, 1999). During the 1800s the new high-sucrose canes discovered in Polynesia supplanted these original hybrid canes. However, though productive and fertile, these cane varieties were susceptible to disease and from the 1920s, they were replaced by modern hybrid cultivars (complex hybrids of *Saccharum cultum* Lloyd Evans and Joshi, *Saccharum officinarum* L. and *Saccharum spontaneum* L. (Lloyd Evans & Joshi, 2016)). As a result, the early history of the production of the first commercial sugarcane hybrids remains obscure, though hybrids generated in Java and Coimbatore, India, predominate in the ancestry of almost all modern sugarcane hybrid cultivars. These new modern hybrids possessed

partly restored fertility, though pollen sterility varies amongst genotype and even in optimal conditions never reaches 100% (Subrananyam & Andal, 1984).

As most sugarcane cultivars were bred during the past 100 years, it has been hard to find a method to reliably characterize the sugarcane breeding population phylogenetically. Though initially promising, chloroplast genomes tend to be highly stable and there are insufficient sequence differences between them to resolve the divergence of close sister cultivars (D Lloyd Evans, unpublished data).

Plant mitochondrial genomes are significantly different from their animal counterparts (Gualberto et al., 2014). Indeed, land plant mitochondrial genomes can vary in size between 187 Kb in *Marchantia polymorpha* L. (Ohya et al., 1986) to 11.3 Mb in *Silene conica* L. (Sloan et al., 2012). However, the mitochondrial genome of the green alga *Chlamydomonas reinhardtii* Dangeard at 15 800bp is the smallest yet assembled (Lister et al., 2003). The plasticity of mitochondrial genomes, leading to genome expansion, arises primarily from repeat sequence, intron expansion and incorporation of plastid and nuclear DNA (Turnel, Otis & Lemieux, 2003; Bullerwell & Gray, 2004). Moreover, plant mitochondria employ distinct and complex RNA metabolic mechanisms that include: transcription; RNA editing; splicing of group I and group II introns; maturation of transcript end and RNA degradation and translation (Hammani & Giege, 2014).

The accumulation of repetitive sequences in plant mitochondrial genomes cause frequent recombination events and dynamic genome rearrangements within a species leading to the generation of multiple circular DNA strands with overlapping sequence and different copy number (Chang et al., 2011; Allen et al., 2007; Guo et al., 2016). In such cases, the complete

genome is referred to as the master circle, with the DNA circles derived from recombination referred to as minicircles (subgenomic circles). Though the current convention is to represent the mitochondrial genome as a single DNA circle (often resulting in duplication of repeat sequence in the final assembly), this is not always noted (Mower et al., 2012).

There are also documented cases where the master circle no longer exists and the genome consists of multiple circular strands of DNA without any shared sequence that could facilitate recombination (e.g. *Silene vulgaris* (Moench) Garacke, *S. noctiflora* L., *S. conica*, *Cucumis sativus* L.) (Sloan et al., 2012; Alverson et al., 2011) Functionally, plant mitochondrial genomes are unlikely to be limited to a single origin of replication (Mackenzie & McIntosh, 2006) (just as in their chloroplast counterparts (Krishnan & Rao, 2009)), though there has been only a single study analyzing in detail the transcription of the plant mitochondrion in *Petunia ×hybrida* hort. ex E. Vilm. (de Haas et al., 1991). The mitogenome can be dynamic, with some plants possessing multipartite maps, typically containing fewer than three chromosomes that can be assembled into circular, linear, branched or sigmoidal forms (Gualberto & Newton, 2017). In contrast, multichromosomal maps can contain tens of linear or circular chromosomes (Sanchez-Puerta et al., 2017).

Break-induced repair and recombination has been proposed as a potential source for mitochondrial genome expansion and could account for the long repeat sequences often found in plant mitochondria (Christensen, 2013). These long repeats, along with DNA shuffling between the nuclear and plastid genomes can confound efforts to assemble plant mitochondrial genomes by introducing branch points within the assembly graph that lead to multiple sequences including mitochondrial, nuclear and chloroplast sequence being incorporated in an

assembly. These phenomena, along with the relatively large size of plant mitochondrial genomes, make them difficult to assemble. However, these effects *in vivo* potentially introduce variable sequences that could be useful in comparing closely related cultivars.

Compared with the chloroplast and nuclear genomes, the mitochondrion is also unusual in that it retains more bacterial-like transcript processing, whereby, in general, transcripts targeted for degradation have poly-A extensions (Gagliardi et al., 2004). Though there may also be a secondary poly-A mechanism protecting stress-induced transcripts (Adamo et al., 2008).

The plant mitochondrion is also typically responsible for a phenomenon known as Cytoplasmic Male Sterility (CMS), a maternally inherited trait that typically results in a failure to produce functional pollen or functional male reproductive organs (Suzuki et al., 2013). The phenomenon of CMS has been reported in over 150 species of flowering plants (Carlsson et al., 2008). The highly recombinogenic, repetitive nature of plant mitogenomes has been linked to CMS and, indeed, CMS is typically conferred via chimeric genes whose generation has been associated with the presence of large repeats (Galtier, 2011). Typically CMS is counteracted by the presence of restorer-of-fertility (*Rf*) genes in the nuclear genome (Huang et al., 2015). Functionally, there are three main routes to CMS in plants: mtDNA recombination and cytonuclear interaction; regulation of CMS transcripts via RNA editing and direct protein interactions whereby CMS protein transmembrane domains directly disrupt or alter the permeability of the mitochondrial outer membrane, thus interfering with energy production (Chen et al., 2017).

Sugarcane mitochondrial chromosomes from a commercial hybrid cultivar SP80-3280 were assembled using Illumina's TruSeq synthetic long reads. This assembly was used as a

template to aid the assembly of the mitochondrial genomes from the cultivars LCP85-384 and RB72343 as well as *Saccharum officinarum* IJ76-514 from New Guinea. Extended annotation of the sugarcane mitochondrial genome revealed a potential cytoplasmic male sterility factor that was a cognate of orf113 previously described in rice (Igarashi et al., 2013).

Transcript reads were mapped to the SP80-3280 mitochondrial chromosomes, revealing the spliceosome of sugarcane mitochondria. Poly-A baited transcripts were mapped to the *Sorghum bicolor* L. cv BTx623 mitochondrion, revealing mitogenomic regions tagged for degradation.

For phylogenetic analyses, Illumina reads from *Saccharum spontaneum* SES234B and *Miscanthus sinensis* cv Andante were partially assembled against the sugarcane template.

We demonstrate the utility of mitochondrial genomes for phylogenetic analyses and show that the sugarcane mitochondrion is transcribed in its entirety and potentially contains a cytoplasmic male sterility factor as well as a functional copy of the chloroplast *rbcL* (RuBisCO large subunit).

Materials and Methods

Sugarcane Mitochondrial Assembly

NCBI was mined for assembled mitochondrial genomes and partial mitochondrial sequences from the genera: *Zea*, *Sorghum*, *Miscanthus* and *Saccharum*. These sequences were used to bait reads from the *Saccharum* hybrid SP80-3280 Illumina's TruSeq synthetic long read dataset (SRA: SRR1763296) (Riaño-Pachón & Mattiello, 2017) using Mirabait 4.9 (Chevreux, Wetter & Suhai, 1999) with a k-mer of 32 and n=50. Baited reads were initially assembled with Cap3, using parameters: -o 1000 -e 200 -p 75 -k 0 (Huang & Madan, 1999). Assembled and unassembled reads were blasted against the initial mitochondrial dataset with an e-value cut-off of $1e^{-9}$ (Camacho et al., 2008). All matching assemblies and reads were added to the read pool and a second round of Mirabait read baiting was performed.

All baited reads were assembled with SPAdes (3.10) (Bankevitch et al., 2012) using default parameters, but with all error correction options enabled. SPAdes contigs were blasted against the mitochondrial dataset and all reads with matches were extracted. These were then blasted against a local collection of *Saccharum* chloroplasts. All assemblies that had almost complete chloroplast coverage were excluded. The final sugarcane mitochondrial assembly pool was baited against the Illumina TruSeq synthetic long read pool using Mirabait again before running a second round of assembly with SPAdes. The process above was repeated twice more.

At this stage, the longest contigs were tested for circularity with Circulator (Hunt et al., 2015). This revealed a complete circular genome of 144639bp. This sequence was labelled as 'potentially complete' and was excluded from further assembly. The remaining contigs were run through four more rounds of baiting and assembly. After these assembly rounds had completed

181 circularity testing with Circulator revealed a second complete chromosome of 300960 bp.

182 Using the two assembled mitochondrial chromosomes of SP80-3280, the mitochondrial
183 genomes of hybrid cultivars LCP85-384 (SRA: SRR427145), RB72454 (SRA: SRR922219)
184 (Grativol et al. 2014) and *S. officinarum* IJ76-514 (SRA: SRR528718) (Berkman et al. 2014)
185 were assembled using a methodology previously developed for chloroplast assembly (Lloyd
186 Evans & Joshi, 2016). Briefly, reads were extracted from the Illumina read pool using Mirabait
187 with a baiting k-mer of 27. These reads were assembled using SPAdes with the SP80-3180
188 mitochondrion employed as an untrusted reference (essentially to resolve repeats). Contigs were
189 scaffolded on the corresponding SP80-3280 mitochondrial assembly and a second round of
190 baiting and assembly was run, this time with a Mirabait k-mer of 31. After a second round of
191 assembly, there were only a small number of short gaps within the assembly. Excising a 2kbp
192 region around the gap and using this for baiting and assembly allowed this completed sequence
193 to fill the gaps. Employing this approach, the two chromosomes of LCP85-384 and RB72454
194 were assembled in their entirety. Chromosomes 1 and 2 of IJ76-514 were partially assembled
195 (both chromosomes contained gaps that could not be closed).

196 Though SRA datasets for *Saccharum* hybrid SP70-1143 existed in GenBank (SRA:
197 SRR952331, SRR871521, SRR871522 and SRR871523) (Grativol et al., 2014), initial assembly
198 using the methods above failed to yield complete mitochondrial chromosomes. To improve
199 coverage, five RNA-seq datasets were downloaded (SRA: SRR1104746, SRR1104748,
200 SRR1104749, SRR619797 and SRR619800) (Bottino et al., 2013; Vargas et al., 2014). These
201 are all single end files and were used as an additional single end dataset (with the --s option) of
202 SPAdes. The combined dataset resulted in a complete hybrid assembly of both SP70-1143
203 mitochondrial chromosomes.

Subsequent to assembly, all assembled mitochondria were finished and polished with a novel pipeline. Raw reads from the SRA pool were mapped back to the assembly with BWA (Li & Durbin, 2009), tagging duplicate sequences with Picard tools (<http://broadinstitute.github.io/picard>), optimizing the read alignment with GATK (McKenna et al., 2010) and finally polishing and finishing with Pilon 1.2.0 (Walker et al., 2014).

Partial Assembly of Related Mitochondria

The mitochondrial genomes of *Saccharum spontaneum* SES234B (SRA: SRR486146) and *Miscanthus sinensis* cv Andante (gifted by CCS, Cambridge, UK) were assembled using the sugarcane SP80-3180 mitochondrial chromosomes and the *Sorghum bicolor* BTx623 (GenBank: NC_008360.1) mitochondrion as templates. Assembled contigs were run through four rounds of baiting with Mirabait (k=31) and assembly with SPAdes. At the same time, reads were mapped to the sugarcane mitochondrial genomes and the Sorghum mitochondrial assembly with BWA (Li & Durbin, 2009). Assemblies and mappings from *Saccharum spontaneum* SES234B and *Miscanthus sinensis* cv Andante, along with the Sorghum mitochondrial assembly were mapped to the sugarcane mitochondrial chromosomes using BLAST. These mappings were employed for all subsequent phylogenetic analyses.

Mitochondrial Annotation

Open Reading Frames (ORFs) were initially predicted using Open Reading Frame Finder (<https://www.ncbi.nlm.nih.gov/gorf/gorf.html>). All tRNA genes were identified using tRNAscan-SE (Schattner et al., 2005). In addition, genes and exons were extracted from the existing *Sorghum* and *Zea mays* L. mitochondrial entries in GenBank. These features were mapped to the SP80-3280 assemblies using Exonerate 2.2.0 (Slater & Briney, 2005). A custom

BioPerl script extracted the Exonerate mapped features and compared them with predicted ORFs to determine confirmed genes. These genes were further checked with the plant mitochondrial genome annotation program Mitofy (Alverson et al., 2010). Repeats were identified using REPuter v3.0 (Kurtz et al., 2012) along with self-blasting the mitochondrial chromosomes to themselves. For chloroplast genes and other features, all genes and features were extracted from the chloroplast genome of sugarcane cultivar RB72454 (NCBI: LN849914) as well as the mitogenomes of *Oryza rufipogon* Griff. strain RT98C (NCBI: BAN67491) (Igarashi et al., 2013) and the *Oryza sativa* L. Indica cv Hassawi mitochondrion (NCBI: JN861111) (Zhang et al., 2012). Features were mapped with blast and manually added to the SP80-3280 mitochondrial annotation files. The high quality annotation of the SP80-3280 mitochondrial genomes was used as the basis for mapping features to the LCP85-384, RB72454 and SP70-1143 assemblies using the Rapid Annotation Transfer Tool (RATT) (Otto et al., 2011). Completed and annotated mitochondrial assemblies were deposited in ENA under the project identifier PRJEB26367. The partial assembly of the IJ76-514 and the hybrid assembly of the SP70-1143 mitochondrion were deposited in the Dryad digital repository (<doi available upon acceptance>).

Sugarcane Chloroplast Assembly

The chloroplast of *Saccharum* hybrid cultivar SP70-1143 was assembled from NCBI sequence read archive datasets as well as transcriptomic datasets, as described previously (Lloyd Evans & Joshi, 2016). The SP80-3280 chloroplast was assembled from TruSeq synthetic long reads, using our standard assembly pipeline, except for the following changes in Mirabait parameters: -k 32 -n 150. The SP80-3280 chloroplast was also assembled from transcriptomic data (SRA: SRR1979660 and SRR1979664) (Mattiello et al., 2015). Transcriptomic assembly

resulted in six contigs covering all the chloroplast apart from the ribosomal RNA region, where there were 26 overlapping contigs. GC content (Supplemental Table S1) was used to identify contigs derived from the chloroplast (GC content = 38.4%), which were made contiguous with CAP3 prior to integration into the main assembly. Assemblies were finished and polished as described for mitochondrial assemblies. The SP70-1143 short read assembly and SP80-3280 TruSeq synthetic long read assemblies were deposited in the ENA under the project identifier PRJEB26685. The EMBL flatfiles corresponding to the transcriptomic assembly of SP80-3280 and the transcriptomic assembly of SP70-1143 can be obtained from Dryad (<doi available upon acceptance>).

Potential *Rf* Transcript Assembly and Sequencing

Restorer of Function (Rf) transcripts were identified from the *Oryza* literature. Orthologues of these genes were identified using the Ensembl Orthology (compara) interface (Viella et al., 2009) or by Phytozome (Goodstein et al., 2011) Blast analysis against the *Miscanthus sinensis* genome assembly (*Miscanthus sinensis* v7.1 DOE-JGI, <http://phytozome.jgi.doe.gov/>). Transcripts and genes were assembled using a bait and assemble strategy (Lloyd Evans & Joshi, 2017) against the SP80-3280 short read transcriptomic and TruSeq Synthetic Long Read genomic datasets. Primers were designed (Table 1) to amplify as much of the transcript sequence as possible (as such the primers were necessarily sub-optimal and could amplify multiple targets). Amplicons were concatenated and sequenced with Oxford Nanopore Technologies MinION prior to assembly with CANU (Koren et al., 2017), as described previously (Lloyd Evans, 2019). Sequences for three transcripts sequenced for N22 and SP80-3280 have been deposited in ENA under the project identifier PRJEB26689.

271

272 **Transcriptomic Data Mapping**

273 Transcriptomic short read datasets (from the high depth SP80-3280 dataset SRA project:
 274 PRJNA244522 (15 datasets) (Mattiello et al., 2015), a pooled cultivar dataset SRA:
 275 SRR849062 (though containing SP80-3280 reads), a pooled tissue dataset SRA: SRR1974519
 276 and a leaf dataset SRA: SRR400035) were mapped to the SP80-3280 mitochondrial
 277 chromosome assemblies and the new SP80-3280 chloroplast assembly using BWA for
 278 unprocessed transcripts and HISAT2 (2.1.0) for spliced transcripts (Kim, Langmead & Salzberg,
 279 2015). All mappings in SAM format were merged with SAMtools (Li et al., 2009) prior to
 280 conversion to BAM and duplicate sequence removal with PICARD and SAMtools prior to
 281 import into IGV (Integrative Genomics Viewer) (Thorvaldsdóttir et al., 2013). The consensus
 282 sequence was exported from IGV, which was also employed to check for non-canonical start
 283 codons and RNA-editing. Transcript counts at each base for the SP80-3280 data were exported
 284 with the SAMtools ‘depth’ command prior to conversion to \log_{10} and drawing on the
 285 mitochondrial genome with Abscissa (Brühl, 2015).

286 For splieosomal analysis and polyA baited read analyses SP80-3280 transcriptomic
 287 reads were mapped to the SP80-3280 mitochondrial chromosomes and *Sorghum biocolor*
 288 BTx623 polyA baited transcriptomic reads (SRA: ERR2097035; ERR2097063; ERR2097067;
 289 ERR3063529 and ERR3087932) were mapped to the *Sorghum bicolor* mitochondrion
 290 (GenBank: NC_008360.1) initially with BWA. In all cases paired end reads were used and reads
 291 where the mate did not map correctly or within the correct distance were excluded from further
 292 analyses as these could represent genomic contamination. From the total mapped read pool,

reads only mapping to the forward strand were extracted with the **Samtools** (Li et al., 2009) command “samtools view -F 20 <bam-file> > se-reads.sam”.

Reads were converted back to fastq format and were re-mapped to the respective genomes with HISAT2 (Kim, Langmead & Salzberg, 2015), a fast read mapper that allows for long indels. Mapped files were converted to BAM format with **samtools** and were imported into the IGV viewer (Thorvaldsdóttir et al., 2013) for further analyses

Phylogenetic Analyses

Assemblies of sugarcane mitochondrial chromosome 1 and chromosome 2 along with mappings of *S. officinarum* chromosome 1 and partial chromosome 2 and *Miscanthus*, *S. spontaneum* and *Sorghum bicolor* assemblies and contigs mapped to sugarcane mitochondrial chromosomes were aligned with SATÉ 2.2.2 (Liu et al., 2009) using default options and the GTRGAMMA model, prior to manual correction of the assembly. Missing sequence was represented by Ns. Regions of the assembly with over 20nt represented by a single sequence only were trimmed down to 10bp to reduce long branch issues. Chromosome 1 alignments and Chromosome 2 alignments were merged with a custom Perl script. Independent analyses were performed on the chromosome 1 dataset, chromosome 2 dataset and the merged dataset. In all cases, the assemblies were partitioned into coding gene, tRNA + rRNA and non-coding partitions. Partition analyses with jModelTest2 (Darriba et al. 2012) revealed GTR+ Γ to be an acceptable model for all partitions.

To determine the best topology, two independent partitioned runs of RAxML (version 8.1.17) (Stamakis, 2006), using different seeds, were run with 100 replicates. Both runs yielded the same best tree topology and this was used as the reference for all future analyses.

Concatenated trees were reconstructed using both maximum likelihood (ML) and Bayesian approaches and rooted on *Sorghum bicolor*. The ML tree was estimated with RAxML using the GTR + Γ model for all 5 partitions, and 6000 bootstrap replicates. The Bayesian tree was estimated using MrBayes v.3.2.1 (Ronquist and Huelsenbeck 2003) using a gamma model with six discrete categories and partitions unlinked. Two independent runs with 25 million generations each (each with four chains, three heated and one cold) were sampled every 1,000 generations. Convergence of the separate runs was verified using AWTY (Nylander et al., 2008). The first six million generations were discarded as burn-in. The ML trees and the MB trees were mapped onto the best topology from the initial RAxML run with the SumTrees 4.0.0 script of the Dendropy 4.0.2 package (Sukumaran & Holder, 2010).

Due to the large size of the combined (chr1+chr2) and chromosome 1 datasets, divergence times on the smaller chromosome 2 alignment only were estimated using BEAST 2.4.4 (Drummond et al., 2012), on an 18-core server running Fedora 25, using four unlinked partitions (as above). However, as chromosome 1 and the combined partition gave the same tree topology, divergence times would not be expected to vary between datasets. The analysis was run for 50 million generations sampling every 1,000th iteration under the GTR + Γ model with six gamma categories. The tree prior used the birth-death with incomplete sampling model (Drummond et al., 2012), with the starting tree being estimated using unweighted pair group method with arithmetic mean (UPGMA). The site model followed an uncorrelated lognormal relaxed clock (Drummond et al. 2006). The analysis was rooted to *Sorghum bicolor*, with the divergence of *Sorghum* estimated as a normal distribution describing an age of 7.2 ± 2 million y (Lloyd Evans & Joshi, 2016). Convergence statistics were estimated using Tracer v.1.5 (Rambaut et al., 2013) after a burn-in of 15,000 sampled generations. Chain convergence was estimated to have been

met when the effective sample size was greater than 200 for all statistics. Ultimately, 30,000 trees were used in SumTrees to produce the support values on the most likely tree (as determined above) and to determine the 95% highest posterior density (HPD) for each node. All final trees were drawn using FigTree v.1.4.0 (<http://tree.bio.ed.ac.uk/software/figtree/>) prior to finishing in Adobe Illustrator. Final alignments and phylogenetic trees are available from the Dryad digital repository (<DOI available on acceptance>).

Mitochondrial and Chloroplast Comparisons

Mitochondrial and chloroplast chromosome comparisons (within and between sugarcane cultivars) were performed with NCBI Blast (Altschul et al., 1990), Mauve (Darling et al., 2004) and EMBOSS Stretcher (Rice, Longden & Bleasby, 2000). EMBOSS Stretcher output was analyzed with a custom Perl script to detect and quantify substitutions, insertions and deletions between the two genomes.

GC Content Analyses

GC content varies between the chloroplast, mitochondrion and the nuclear genome. We used our assemblies to compare GC content between related mitochondria, related chloroplasts, the assembled genomes of Sorghum and maize and the synthetic long read pool of sugarcane (excluding mitochondrial and plastome reads) using the EMBOSS cusp application. The data obtained from this study was used to ensure that our mitochondrial assembly arose only from mitochondrial data and to examine introgression of sequence from the chloroplast and nuclear genome into the mitochondrion of sugarcane.

Transposable Element Analyses

The presence of transposable elements within the sugarcane mitochondrion was examined using

the Poaceae database as query for the Genetic Information Research Institute
(<http://www.girinst.org/censor/index.php>) Censor application (Kohany et al., 2006).

Molecular Modelling of rbcL

The protein sequences of sugarcane chloroplast rbcL (RuBisCO large subunit) and mitochondrial rbcL were submitted to the Phyre² server (Kelley et al.; 2015) for homology modelling. PDB files from Phyre² intensive modelling were downloaded and prepared for molecular dynamics (MD) simulation using the Protein Preparation Wizard of the Maestro molecular modelling software (v.9.6; Schrödinger, Inc.). The model included all hydrogen atoms from the start, but the polar interactions of the *His* residues were manually checked and the protonation states selected to optimize the hydrogen bond network.

MD simulations were performed to confirm that the 3D structure was stable without unfolding or any significant changes in secondary structure. The Groningen Machine for Chemical Simulations (GROMACS) (Abraham et al. 2015) with the CHARMM force field was employed for this purpose and solvated our model in a cubic box with TIP3P water. The system was charge equilibrated with 8 sodium ions before being energy minimized. After energy minimization, the systems were equilibrated by position restrained molecular dynamics at constant temperature of 300 K and a constant pressure of 1 atm for about 100 ps before running a 200 ns molecular dynamics simulation using the CHARMM force-field. The final models were compared with each other and with the original spinach template to ensure conformational stability.

Final models were imported into USCS Chimera (Pettersen et al., 2004) and were superimposed with the MatchMaker tool and RMSD differences were determined from the Reply Log panel.

381

Results

Mitochondrial Genome Assembly and Annotation

An iterative approach was used to assemble the mitochondrial genome of *Saccharum* hybrid SP80-3280 using Illumina's TruSeq synthetic long reads. This resulted in the assembly of two mitochondrial chromosomes: one of 144639 bp and one of 300960 bp (Figure 1). Average read depth on both chromosomes was 12.4x. No reads were found that linked the two chromosomes, indicating that the sugarcane mitochondrion exists as two separate chromosomes without a maxicircle. Whilst there were a large number of repeats within each chromosome, few repeats were found to be common between both chromosomes. This makes it unlikely that the chromosomes can recombine to form a master circle.

In addition, the mitochondrial genomes of LCP85-384 and RB72454 were assembled. We also attempted assembly of the mitogenome of cultivar SP70-1143. There was insufficient coverage from nuclear sequence to completely assemble the two mitochondrial chromosomes. As a result, a hybrid approach was attempted, adding five RNA-seq datasets to improve overall coverage. This resulted in the complete assembly of the two SP70-1143 mitochondrial chromosomes.

All mitochondrial genomes had a 15 Kb direct repeat sequence and a 4 Kb inverted repeat on chromosome 1 (Figure 1). Full annotation of the genomes (based on previous mitochondrial annotations, mapping chloroplast genes and mapping additional genes from rice and maize mitogenomes) revealed 72 unique open reading frames plus 26 duplicate copies, 14 complete chloroplast genes and 27 partial chloroplast gene fragments. Of these, 64 genes are encoded by a single exon and eight genes are encoded across multiple exons. Moreover, trans-

splicing of group II introns were observed in three genes: *nad1*, *nad2* and *nad5*. The genes *nad2* and *nad5* have exons split between chromosome 1 and chromosome 2 (a similar phenomenon is seen in *Silene vulgaris* (Sloan et al., 2012)). Sugarcane mitochondrial genomes had the same gene content as sorghum, with the exception of *trnL-CAA* and *rbcL-cp*, which are present in sugarcane, but absent from sorghum.

Comparisons of the mitochondrial assembly of SP80-3280 with the chloroplast genome assembly from the same cultivar revealed that seven of the total tRNA genes plus 14 other genes were derived from the chloroplast genome, mostly present in large sections of transferred DNA.

Assembly of the *Saccharum spontaneum* SES234B mitochondrial genome was attempted. Large contigs were obtained, demonstrating considerable sequence conservation with the sugarcane hybrid assemblies. Examining the assembly graphs for the *S. spontaneum* cv SES234B mitogenome revealed that there were multiple reads joining chromosome 1 and chromosome 2 as based on the sugarcane hybrid assemblies. This indicates that either *S. spontaneum* mitochondrion exists as a single circular genome or there is a different organization of this species' mitochondrial chromosomes. As a result, we were not able to completely assemble the mitogenome of *S. spontaneum*. As a compromise, the assembled *S. spontaneum* mitochondrial contigs were mapped to the sugarcane chromosome 1 and chromosome 2 assemblies. This mapping was subsequently used for phylogenetic analyses.

An attempt at assembling reads for *Miscanthus sinensis* cv Andante revealed a similar pattern to that of *S. spontaneum*, again indicating that the mitochondrion of this species also exists as a single chromosome. Again, *M. sinensis* contigs were mapped to the sugarcane chromosome 1 and chromosome 2 assemblies for subsequent use in phylogenetic analyses.

Assembly of *S. officinarum* cv IJ76-514 was attempted, using our previous chloroplast assembly for this cultivar (Lloyd Evans & Joshi, 2016) all reads mapping to the chloroplast were removed with BWA and SAMtools. The remaining reads were baited and assembled based on the SP80-3280 mitochondrial genome assembly. It took five rounds of baiting and assembly to fully assembly chromosome 1 (apart from six small gaps), but after 10 rounds chromosome 2 still had significant gaps. This could mean low coverage of certain genomic regions, but it also indicates more sequence variations than had previously been reported.

Transposable Element Analysis

Censor (Kohany et al., 2006) analyses revealed 114 potential transposable elements in chromosome 1 and 48 potential transposable elements in chromosome 2. The coordinates of the transposable elements in chromosome 1 and chromosome 2 of the SP80-3280 mitochondrial genome are given in Supplemental Table S2. Though there are many fragments of transposons within the mitochondrial genome, none are functional and all are degraded from their original genomic ancestors.

Phylogenomic Analyses

BLAST analysis of our assembled SP80-3280 mitochondrial chromosomes against the assembled mitochondrial genome of *Sorghum bicolor* BTx623, revealed that 345 kb of its 468 kb genome is represented in our assembly, although, substantially rearranged. Thus, a considerable portion of the total mitochondrial repeat sequences are shared between the two species. This includes 3 kb of the inverted repeat and the entire 15kb direct repeat, though split into two parts in *Sorghum*, with the entire repeat existing as only a single copy in the *Sorghum* mitogenome. This indicates that our strategy of mapping assembled contigs from *Miscanthus*

and *Saccharum spontaneum* onto the sugarcane assembly is valid and results in accurate sequence for phylogenetic analyses.

Mitochondrial chromosome assemblies of the sugarcane hybrids: SP80-3280, Khon Kaen 3, LCP85-384, RB72343, and SP70-1143 were separately aligned to the two chromosomes from *S. officinarum* IJ76-514 as well as the mapping of the Sorghum mitogenome to the two sugarcane chloroplasts and the mappings of *S. spontaneum* and *M. sinensis* contigs to the two chromosomes of sugarcane. Each chromosome was aligned independently, prior to both alignments being merged.

Maximum likelihood analyses of chromosome 1, chromosome 2 and the combined dataset revealed exactly the same tree topology (Figure 2). The chromosome 1 and chromosome 2 alignments were taken further for ML bootstrap and BI support determination. Both analyses revealed 100% support for all branches. The data for chromosome 2 only is shown in Figure 2, as only this dataset was employed for BEAST analyses to generate a chronogram. The phylogeny shows the expected topology and is consistent with our previous studies (Lloyd Evans & Joshi, 2016; Lloyd Evans, Joshi & Wang, 2019). We also clearly see the relationships between SP70-1143, SP80-3280 and Khon Kaen 3.

Transcriptomic Read Mapping

High depth RNA-seq data were available for sugarcane cultivar SP80-3280 and were mapped to the mitochondrial genome for spliceosome analysis. Unfortunately, there were insufficient polyA-baited reads to allow mapping to the sugarcane mitochondrion. As a

result mapping of polyA-baited reads was performed to the *Sorghum bicolor* BTx623 mitochondrial genome instead.

After pre-processing to ensure both reads of paired end data mapped to the appropriate mitochondrial genome, reads were converted to forward strand only. These reads were re-mapped with HISAT2 and imported into IGV prior to analysis.

Transcript mapping to the SP80-3280 mitochondrial genome chromosomes revealed a complex pattern of splicing events, many spanning the two chromosomes (Figure 3). The most common splicing event joined the start of chromosome 1 with the start of chromosome 2. Internally splicing events were from one locus hotspot to another locus hotspot that spanned a few hundred to a few thousand bases. Thus splicing events were not targeted to a few bases as is typical in eukaryotic genomes. In addition, of 222 splicing events (only counting splice sites with ≥ 10 reads mapped) 110 (49.55) were inside coding sequences — which is almost half — an unexpectedly high number. The full analyses of splice sites in the SP80-3280 mitochondrial chromosomes is given in Supplemental Table S3.

Compared with other plant genomes, mitochondria are unusual in that they retain much (though not all) of their α -proteobacterial antecedents' processing (Gagliardi et al., 2004). Indeed, under non-stressed conditions mitochondria add poly-A tails to those transcripts marked for degradation. To examine this process polyA-baited reads were mapped to the BTx623 mitochondrial genome. As can be seen from Figure 4, polyA baited reads map to distinct 'islands' within the sorghum mitochondrial genome. Examining these islands, of the 35 identified, only five contained genes annotated in the Sorghum

mitogenome. However, when the chloroplast and nuclear genomes were included in searches along with mitochondrial gene duplications an additional 24 genes were identified. The remaining polyA tailed regions were all repeat regions, intronic regions and intragenic regions. The full analysis of polyA read islands mapped to genes is provided in Supplemental Table S4.

Chloroplast Assembly and Analyses

The currently published SP80-3280 chloroplast was assembled in 2002 (Calsa Jr et al., 2004). The state of the art in terms of chloroplast assembly and sequence finishing has moved on considerably during the intervening decade and a half. We re-assembled the SP80-3280 chloroplast from Illumina's TruSeq synthetic long reads, using our novel sequence-finishing pipeline for assembly polishing. Analyses showed that our assembly differed from the GenBank accession by only 8 substitutions and a single insertion. To see if this was typical or unusual, we also assembled the SP80-3280 chloroplast from transcriptomic data, as well as assembling the chloroplast of the closely related cultivar SP70-1143. Comparisons were also made to the LCP95-384, RB72454 and Q165 sugarcane chloroplasts that we had previously assembled (Lloyd Evans & Joshi, 2016), as well as the Q155 (GenBank: NC_029221) (Hoang et al., 2016), NCo310 (GenBank: NC_006084) (Asano et al., 2004) and RB867515 (GenBank: KX507245) (Barbosa et al. 2016) assemblies from GenBank.

Transcriptomic Coverage of Multiple-chromosome Mitogenomes

Mapping of transcriptomic data from **eighteen sugarcane RNA-seq** datasets to the SP80-3280 assembly revealed that the mitogenome of sugarcane is completely transcribed (Figure 5). We observed a mix of processed (spliced) and unspliced transcripts, with 99.995% of the

mitochondrial chromosomes covered by sequence (i.e. not Ns). Only in a single instance, were all mapped transcripts processed. This being the start codon of *nadl*, where the entire set of DNA reads had cytosine in the first position, whilst all the RNA-seq reads had an Uracil (see Figure 5 for the mapping data). Moreover, there was complete coverage of the SP80-3280 chloroplast by transcriptomic data.

We also assembled the sugarcane SP80-3280 chloroplast from transcriptomic data. The assembly was the same length as our genomic assembly (Table 2). However, there were 45 sequence substitutions.

Molecular Modelling of Sugarcane *rbcL*

Annotation of the sugarcane mitochondrion revealed that sugarcane might, uniquely, possess a functional *rbcL* molecule in its mitochondrion. The C-terminus of this is different from that of the chloroplast model, but a new stop codon is in frame and the altered amino acids are all within the disordered C-terminus and do not contribute to the functional core of the molecule. To see if this mitochondrial copy of *rbcL* might be functional, the protein sequences of the chloroplast and mitochondrial copies of sugarcane *rbcL* were modelled by homology with the Phyre² server. In both cases, >97% of all residues were modelled with 93% confidence. The template for modelling was non-activated spinach **rubisco in** complex with its substrate: ribulose-1,5-bisphosphate (PDB: 1RCX) (Taylor & Anderson, 1997). To ensure that the initial mapping had not over-constrained the molecules to the same structure molecular dynamics simulations were performed. Superimposing the sugarcane *rbcL* structures onto the spinach template revealed that all contacts made by spinach *rbcL* with the substrate are also made by the sugarcane chloroplastic and mitochondrial versions of the *rbcL* subunit, indicating that sugarcane mitochondrial *rbcL* could be active and functional. In addition, superposition of the sugarcane

536 models revealed that they were essentially identical (Figure 6) with a root mean square
 537 difference (RMSD) of 0.356Å.

Discussion

Mitochondrial Genome Assembly and Annotation

Using Illumina TruSeq Synthetic Long Reads and an iterative approach we were able to assemble the complete mitochondrial genome of sugarcane cultivar SP80-3280. Whilst there were a large number of repeats within each chromosome, few repeats were found to be common between both chromosomes. This makes it unlikely that the chromosomes can recombine to form a master circle.

Subsequent to our initial assembly of the SP80-3280 mitochondrion, the paper of Shearman et al., (2016) was published. This revealed an independent assembly of the mitochondrion of a sugarcane hybrid cultivar (Khon Kaen 3). Their chromosome 1 was 300784 bp long and their chromosome 2 was 144698 bp long. Differences were due to a single deletion in SP80-3280 chromosome 1 and a single insertion in SP80-3280 chromosome 2 (both in AT rich repeat regions). The remainder of the sequence is almost identical. This is hardly surprising, however, as both cultivars share a (recent) common female parent. As we have the complete mitochondrial sequences of SP80-3280, SP70-1143 and Khon Kaen 3, the mitochondrial genome of SP70-1143 was also assembled the relatedness between these three cultivars could be examined.

The mitogenomes of LCP85-384 and RB72454 are more divergent, with chromosome sizes of 300943, 144679 and 300828, 144692, respectively. The main differences being insertions and deletions within AT-rich repeat regions as well as single nucleotide substitutions distributed throughout the genome.

Within the sugarcane chloroplast genome, a single pseudogene, ACR, is conserved from an ancient translocation event with mitochondrial DNA to **the. BLAST** (Altschul et al., 1990) analyses against a local database of whole and partial plastid sequences reveals that this event occurred in the Petrosaviales (about 120 million years ago (Mennes et al., 2013)) long before the loss of ACR in the mitochondria of true grasses.

The assembly of *S. officinarum* IJ76-514 proved to be more interesting. A previous study, using **blast** to map IJ75-514 reads to a sugarcane mitochondrial genome assembly revealed very few differences between the accessions (Shearman et al., 2016). To see if this was the case, we employed a more systematic assembly approach, attempting to assemble the mitogenome of *S. officinarum* IJ76-514 from scratch. Chromosome 1 was assembled with only 6 small gaps, but chromosome 2 had significant gaps that could not be closed. Indeed, base-by-base comparisons of the SP80-3280 assemblies and the IJ76-514 assemblies revealed a total of 1102 sequence variations (Table 2).

Phylogenomic Analyses

The mitochondrial phylogeny (Figure 2) shows the expected topology, with *Sorghum bicolor* as the outgroup. *Miscanthus* is 4.3 million years divergent from sugarcane with *S. spontaneum* 1.37 million years divergent. *Saccharum officinarum* diverged 590 000 years ago from the lineage of sugarcane hybrid cultivars. This confirms our previous findings (Lloyd Evans & Joshi, 2016), demonstrating that the lineage leading to modern sugarcane hybrid cultivars is a separate species (*Saccharum cultum* Lloyd Evans and Joshi) from *Saccharum officinarum*. The dating of the separation of genus *Saccharum* from *Miscanthus* at 4.3 million years and *S. spontaneum* from the other *Saccharum* species at 1.37 million years is in good

agreement (3.8 million years and 1.4 million years) with our previous study (Lloyd Evans & Joshi, 2016).

As expected, SP70-1143 emerges as ancestral to both SP80-3280 and Khon Kaen 3 (Figure 2), confirming the shared parentage of these three cultivars.

The presence of indels and sequence variants within the mitochondrial genomes of sugarcane cultivars, even when they share a recent common female ancestor indicates that mitogenomes could be the sequence of choice for analyzing the relationships between closely related cultivars. However, complete (or very near complete) mitochondrial genomes need to be used for this type of analysis. Potentially, this could work well within the cultivar collection, as they are likely to be closely related sequences and phylogenetic confusion due to cross-over with the nuclear genome will be minimal. The problem comes with obtaining a meaningful outgroup. However, the approach undertaken in this paper of mapping to a reference genome prior to alignment shows a way forward. Indeed, our partial alignment of a *Miscanthus* mitogenome to the sugarcane reference would make an ideal outgroup for such an analysis.

Transcriptomic Read Mapping

For the first time we have mapped genome-scale transcriptomic reads to a complex (multi-chromosome) plant mitochondrial genome. The majority of spliced reads are between the start of mitochondrial chromosome 1 and the end of mitochondrial chromosome 2 (shown boxed in Figure 3). Thus it appears that the two chromosomes of the sugarcane mitochondrial genome are combined at the spliceosomal level. Indeed, of the 111 significant splicing events identified (Supplemental Table S3) 23 (20.7%) were between the two mitochondrial chromosomes. Unlike in eukaryotic genomes, splice sites were clustered at genomic loci (Figure

3, Supplemental Table S3) and almost 50% of splice sites were within coding regions. Recently (Tsuji-mura et al., 2018) reported on the three mitochondrial chromosomes of *Allium cepa* (onion) CMS line Momiji-3. However, unlike in sugarcane the mitochondrial sub-circles of onion can combine into a master circle through recombination at repeats. Though they mapped transcriptomic reads to the mitogenome, they did not report complete expression and they did not perform spliceosomal analyses. They reported only on RNA editing within the genome, describing 635 editing positions.

Unfortunately, there were insufficient polyA baited reads available in NCBI's sequence read archive to analyze the regions that had polyA tails and were programmed for degradation in the sugarcane mitochondrial transcriptome. As a result, polyA baited reads were mapped to the *Sorghum bicolor* BTx623 mitogenome instead. PolyA reads only covered 18.9% of the mitochondrial genome. The regions covered are shown in Figure 4 and full details of the regions and the genome annotation associated with them are given in Supplemental Table S4. In all cases, regions covered are secondary copies of mitochondrial genes, individual exons, pseudogenes, genes captured from the chloroplast, repeat regions, introns and intra-genic regions. These are precisely the regions that would be expected to be tagged for degradation in a mitochondrial genome that is completely transcribed.

Identification of Possible CMS factors in Sugarcane

Mapping of the *Oryza rufipogon* strain RT98C (NCBI: BAN67491) mitochondrial features to the sugarcane mitochondrion revealed unexpected homology between orf113 in *O. rufipogon* and a putative 345nt open reading frame (ORF) in the sugarcane mitogenome (chromosome 1) see Figure 1. Orf113 is labelled as a 'candidate cytoplasmic male sterility gene'

as identified by Igrashi et al. (2013) and which has subsequently been demonstrated to be the causative agent of CMS in the RT98A (without restorer functionality) line of *O. rufipogon* (Toriyama et al., 2013). Typically such CMS factors are gene fusions and contain a transmembrane domain. At the protein level, the *O. rufipogon* and sugarcane ORFs differ by 14 internal amino acid substitutions (five of which are functionally synonymous) and the substitution of IleIle in the rice C-terminus of the protein for TyrLysAsn in the sugarcane orthologue's C-terminus. Both proteins have a predicted transmembrane helix (Figure 4) and both proteins are derived from a nad9 precursor in the mitochondrial genome. Indeed, at the DNA sequence level the CMS protein in sugarcane is identical to nad9 in rice except for seven base substitutions. Interestingly, bases 1 to 249 of the sugarcane mitochondrial protein mapped twice to a sugarcane SP80-3280 genomic sequence (NCBI: MF737055). This potential CMS factor was found in all the modern sugarcane hybrid mitochondrial genomes assembled in this study and was also found to be present (but not annotated) in the previously published Khon Kaen 3 mitochondrial genome (Shearman et al., 2016). As a direct orthologue of rice orf113 it is therefore highly likely that this newly discovered sugarcane mitochondrial open reading frame is a cytoplasmic male sterility factor.

Additional blast analyses revealed that a pseudogene corresponding to orf113 was present in the maize mitochondrial genome but that an orthologue was not present in the *Sorghum bicolor* mitochondrion. The complete CMS sequence was identified in the *S. officinarum* IJ76-514 mitochondrial assembly as well as the mapped assembly of *Miscanthus sinensis* but was not found in a complete and translatable form in the mapped assembly of *S. spontaneum* despite the region that contains this sequence being present in the *S. spontaneum* mitogenome contigs. Indeed, our partial assembly of the *S. spontaneum* mitogenome indicates that the *S. spontaneum*

648 mitochondrion may have undergone re-arrangement only 320bp upstream of the CMS gene
649 locus. This leads to the intriguing possibility that the CMS factor has been lost and re-gained
650 several times through the evolution of the Andropogoneae.

651 The flip-side of cytoplasmic male sterility is that for pollen viability to be re-gained a
652 restorer factor (*Rf*) gene must be present in the nucleus. Studies on rice **revel three** main types of
653 *Rf* genes: Pentatricopeptide-repeat (PPR) proteins (Gaborieau, Brown & Mireau, 2016), ubiquitin
654 domain proteins (Fujii et al., 2014) and glycine-rich proteins (Itabashi et al., 2011). An example
655 of each was taken from characterized *Oryza sativa* proteins (genome references: Rf1:
656 Os05g0207200; Ubiquitin domain containing protein Os10g0542200 and Rf2 Os02g0274000)
657 and the *Sorghum bicolor* BTx623 orthologues were identified. Using these the sugarcane
658 orthologues were assembled using a bait and assemble methodology (Lloyd Evans & Joshi 2017).
659 Single sugarcane orthologues were obtained for the ubiquitin domain and glycine rich proteins,
660 but multiple PPR proteins were assembled. Typically this is a large gene family in plants, often
661 with 600 or more members (numbers from orthology in the Ensembl sorghum and maize genome
662 data). Two criteria seem to limit functional PPR proteins in CMS. They must contain a suitable
663 number of duplicated PPR domains and must be targeted to the mitochondrion (Schmitz-
664 Linneweber & Small, 2008). A pipeline was scripted, whereby as many orthologues of the rice
665 PPR protein were assembled as possible. These orthologues were checked for full length CDSes
666 and the CDSes were translated to protein sequence. The proteins were piped to a local
667 implementation of TPrpred2 (Savojardo et al., 2014) and MU-LOC (Zhang et al., 2018) to check
668 for a mitochondrial transit peptide. Of 239 transcripts assembled, only one had a predicted
669 mitochondrial transit peptide and this was taken on for further analyses and validation by PCR

670 cloning and sequencing. Domain analyses of the CMS protein and the three restorer proteins are
671 shown schematically in Figure 4.

672 **Confirmation of Assembled Transcript Sequences**

673 For the three putative restoration of function (*Rf*) transcripts primers were designed and
674 the transcripts were amplified from N22 and SP80-3280 cDNA libraries prior to Oxford
675 Nanopore Technologies (ONT) MinION sequencing and CANU assembly. Sequences have been
676 deposited in ENA under the project accession: PRJEB31395.

677 **Chloroplast assembly and analyses**

678 Comparisons of the SP80-3280, SP70-1143, LCP95-384, RB72454, Q165, Q155, NCo310
679 and RB867515 chloroplast genomes (sampling the Louisiana, Brazilian, Barbadian, Australian
680 and South African breeding programmes) revealed that chloroplast assemblies were essentially
681 identical, with only a few sequence substitutions and insertions/deletions distinguishing
682 chloroplasts from diverse global populations (Table 2). The IJ76-514 chloroplast emerges as an
683 outlier, with 26 substitutions, 2 insertions and 5 deletions as compared with the SP80-3280
684 chloroplast. This shows that modern sugarcane hybrids are derived from a very limited number
685 of female parents, and the chloroplast genomes are almost clonal. The *S. officinarum* IJ76-514
686 chloroplast is more divergent, supporting the evolutionary separation of *S. officinarum* from the
687 modern sugarcane hybrid cultivars.

688 **Sugarcane Mitochondrial *rbcL* Analysis and Modelling**

689 Annotation of the sugarcane mitochondrion revealed a potentially functional copy of the
690 chloroplast *rbcL* molecule. Molecular modelling revealed that despite containing a modified
691 carboxy terminal the second copy of *rbcL* in the mitochondrion of sugarcane was potentially

active. Capture of *rbcL* sequences has previously been demonstrated in the Andropogoneae, However, in previous cases where this phenomenon has been noted the *rbcL* gene has been rendered inactive due to internal frameshifts (Clifton et al., 2004). This is the first instance where a potentially functional *rbcL* molecule has been reported in a grass mitochondrial genome. This could be associated with a relatively recent recombination between the mitochondrial and chloroplast genomes in sugarcane.

Transcriptomic Coverage of Multiple-chromosome Mitogenomes

Mapping of transcriptomic data to the SP80-3280 assembly revealed that the mitogenome of sugarcane is completely transcribed (Figure 5). It was only recently (Shi et al., 2016; Lima & Smith, 2017) that plant chloroplast genomes and a subset of plant mitochondrial genomes were shown to be fully transcribed, and our findings represent the first report of the full transcription of a multi-chromosomal plant mitochondrial genome.

SP80-3280 mitochondrial chromosome 1 had 19 unassigned bases (Ns) divided between four distinct regions of the genome. Mitochondrial chromosome 2 had three unassigned bases divided between three distinct regions of the genome. The chloroplast was 100% covered by transcriptomic reads. As a result, we are confident in saying that the complete plastome complement of sugarcane is transcribed in its entirety.

The SP70-1134 mitochondrial genome, which was assembled from a mix of genomic and transcriptomic data, showed considerable identity to both the Khon Kaen 3 and SP80-3280 genomes (to which it is an ancestor). Comparison with SP80-3280 revealed a total of 118 substitutions in chromosome 1 (of which 55 were compatible with C→U substitutions characteristic of RNA editing). Chromosome 2 revealed 44 substitutions, 22 of which were

714 consistent with RNA editing.

715 Though it has been demonstrated previously (on a small sample) that relatively small
 716 mitogenomes are transcribed in their entirety (Lima & Smith, 2017) this is the first report of the
 717 complete transcription of a multi-chromosomal mitogenome. To demonstrate that the
 718 phenomenon is universal, transcriptomic short reads were also mapped to the multi-partite
 719 mitogenomes of *Silene vulgaris* (7 Chromosomes), *Cucumis sativus* (7 Chromosomes) and
 720 *Allium cepa* L. (2 chromosomes). In all cases, even for mitochondrial chromosomes with no
 721 coding sequences, there was a minimum of 91.74% coverage (Supplemental Document S1).

722 When the transcriptomic assembly of the sugarcane SP80-3280 chloroplast was analyzed
 723 on a single base level, 22 of these substitutions proved to be C→U, characteristic of RNA
 724 editing. The remainder of the substitutions were G→A, indicating a second form of RNA
 725 editing not previously described for chloroplasts. As a result, there were no sequence
 726 differences between the transcript-assembled and the genome-assembled chloroplasts of SP80-
 727 3280 that could not be accounted for by RNA editing. This also adds *Saccharum* hybrids to the
 728 list of plants with chloroplasts that have been demonstrated to be transcribed in their entirety.

Conclusion

We have assembled three sugarcane cultivar mitochondrial genomes from Illumina genomic data. Mapping of transcriptomic RNA-seq reads to the SP80-3280 mitochondrial genome assembly revealed, for the first time, that the complete complex mitochondrial genomes of this plant species are transcribed in its entirety, even when those mitogenomes are subdivided into multiple chromosomes. Mapping of RNA-seq data to the sugarcane mitochondrial genomes revealed multiple splice events, with the major splice species joining chromosomes 1 and 2 together. Thus the two chromosomes of the sugarcane mitochondrion appear to be joined at the transcript and not the DNA level. Interestingly, splice sites seem to be distributed into spliceosomal ‘hotspots’ with many of these occurring in coding sequences. Moreover, the sugarcane mitochondrion may be unique amongst plant mitochondria analysed to date in that there are no signs of repeat or shared sequence between the two mitochondrial chromosomes that would allow recombination into a master circle. Thus, despite only having two chromosomes the sugarcane mitochondrion does not fit into the multipartite map model. Rather, the sugarcane mitochondrion appears to be truly multichromosomal (though with only two chromosomes) and these chromosomes are integrated at the RNA splicing stage.

Unfortunately, there were insufficient polyA-baited transcriptomic datasets available for mapping to the sugarcane cultivar SP80-3280’s mitochondrial genome. As a result polyA-baited reads from the BTx623 cultivar of *Sorghum* were mapped to the corresponding *Sorghum* mitogenome. In all cases, the regions of the sorghum mitogenome covered by polyA reads are exactly those regions that would be expected to be marked for degradation. This confirms the major bacterial-like role of polyadenylation in the mitochondrion — that eradicating unwanted transcripts or non-functional by-products of transcript editing.

Attempts at assembling the mitochondrial genomes of *Miscanthus sinensis*, *Saccharum spontaneum* and *Saccharum officinarum* yielded incomplete assemblies demonstrating that sugarcane hybrids have diverged significantly from all these species. Indeed, when the assembled reads from these species were mapped to the sugarcane mitochondrial chromosome assembly we were able to use them to perform a phylogenetic analysis, which revealed the sister relationship of *Miscanthus* to genus *Saccharum*, *Saccharum spontaneum* to the crown *Saccharum* species/cultivars and *Saccharum officinarum* to *Saccharum cultum* (the female ancestor of modern sugarcane hybrids).

Sequence level analysis of mitogenomes and chloroplast genomes revealed greater variability in the mitogenome, indicating that mitochondrial genomes will be of greater utility in determining the relationships of sugarcane cultivars to each other than chloroplast genomes. Indeed, the lack of variability amongst chloroplast genomes indicates that modern sugarcane hybrids arose from a very small pool of *S. cultum* cytoplasmic donors. Mitochondrial analysis also confirms *S. cultum* as being distinct from *S. officinarum*, adding credence to our previous study (Lloyd Evans & Joshi, 2016).

GC content analysis reveals substantial differences between mitochondrial, plastid and nuclear genome GC contents, meaning that GC content is a viable methodology to distinguish between the three genome types. This is important, as both chloroplast and mitochondrial genomes are transcribed in their entirety, thus it is possible to assemble these plastomes from transcriptomic data (as we have done for both SP80-3280 and SP70-1143 in this study). We also demonstrate that a combination of genomic and transcriptomic data can be used to assemble mitochondrial genomes (as we have done for the *Saccharum* hybrid cultivar SP70-1143).

774 However, without a template, plant mitochondrial genomes remain hard to assemble,
775 though we demonstrate the utility of Illumina's TruSeq synthetic long read technology in
776 mitogenome assembly pipelines.

777 For the first time we demonstrate that sugarcane possesses all the necessary machinery for
778 cytoplasmic male sterility (CMS), including a CMS gene in the mitochondrial genome and
779 representatives of the three main restorer-of-function (*Rf*) genes in the nuclear genome. The
780 homology between orf113 in *O. rufipogon* and the potential CMS factor in the sugarcane
781 mitochondrion with **nad9** suggests that this CMS factor may act by affecting complex I (NADH
782 dehydrogenase) of the electron transfer pathway (Chen et al., 2017). This goes some way to
783 explaining the phenomenon of incomplete pollen infertility in sugarcane and indicates that CMS
784 in sugarcane is only partially restored. These findings also point the way to generating CMS and
785 restorer lines from sugarcane cultivars, which would be a major leap **for ward** for sugarcane
786 breeding.

787

788Funding

789This work was partly funded by the South African Sugarcane Research Institute.

790

791Author Contributions

792DLIE and SVJ conceived the chloroplast assembly from genomic data component; DLIE conceived
793the remaining experimental aspects, designed and performed the experiments. TH performed an
794initial feature mapping to mitochondrial genomes and DMRP provided the SP80-3280 genomic
795and transcriptomic sequence data. SVJ and DLIE supervised TH. DLIE developed all software
796scripts, performed all the analyses, analysed and interpreted the data and wrote the paper. SVJ and
797DMRP critically proofread the final draft of the manuscript. All authors reviewed and accepted the
798final manuscript.

799

800Acknowledgements

801We thank CCS, Waterbeach, Cambridge, for providing the *Miscanthus sinensis* cv Andante
802sequence data and performing the sequencing. We are grateful to Oxford Nanopore Technologies
803for support through their community access programme. We would also like to thank Dr L Ramnath
804for the N22 cDNA library and The British Association of Sugar Technologists for the SP80-3280
805cDNA library.

806

807Data Availability

808 All finished assemblies from this study have been deposited in ENA under the project identifiers
 809 PRJEB26367 (for mitochondria), PRJEB26685 (for chloroplasts) and PRJEB31395 for sugarcane
 810 gene/transcript datasets. Partial assemblies and assemblies based on transcriptomic data or hybrid
 811 data along with all alignments and phylogenetic trees (including partial assemblies) were deposited
 812 in the Dryad Digital Repository (<DOI available on acceptance>). Computer code developed for
 813 this project is available from GitHub: <https://github.com/gwydion1/bifo-scripts.git>.

814

815 **Competing Interests**

816 Declarations of interest: none

References

- Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E. 2015. CROMACS: High performance molecular simulations through multi-level parallelism from laptop to super computers. *Software* X1:19-25.
- Adamo A, Pinney JW, Kunova A, Westhead DR, Meyer P. 2008. Heat stress enhances the accumulation of polyadenylated mitochondrial transcripts in *Arabidopsis thaliana*. *PLoS One*, 3:p.e2889.
- Allen JO, Fauron CM, Minx P, Roark L, Oddiraju S, Lin GN, Meyer L, Sun H, Kim K, Wang C, Du, F. 2007. Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. *Genetics*, 177:1173–1192.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403-410.
- Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD. 2010. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Molecular Biology and Evolution*, 27:1436–1448.
- Alverson AJ, Rice DW, Dickinson S, Barry K, Palmer JD. 2011. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. *The Plant Cell*, 23:2499–2513.
- Artschwager E, Brandes EW. 1958. Sugarcane (*Saccharum officinarum* L.) Agriculture Handbook No 122. United States Department of Agriculture. Washington D.C.

837 Asano T, Tsudzuki T, Takahashi S, Shimada H, Kadowaki KI. 2004. Complete nucleotide
838 sequence of the sugarcane (*Saccharum officinarum*) chloroplast genome: a comparative analysis
839 of four monocot chloroplast genomes. *DNA Research*, 11:93-99.

840 Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
841 Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV. 2012. SPAdes: A New Genome Assembly
842 Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational*
843 *Biology*, 19:455–477. doi:10.1089/cmb.2012.0021

844 Barbosa MHP, Vidigal PMP, Coelho ASG, Peternelli LA, Novaes E. 2016. Complete chloroplast
845 genome sequence and annotation of the *Saccharum* hybrid cultivar RB867515. *Genome*
846 *Announcements*. 4:e01157-16

847 Berkman PJ, Bundock PC, Casu RE, Henry RJ, Rae AL, Aitken KS. 2014. A survey sequence
848 comparison of *Saccharum* genotypes reveals allelic diversity differences. *Tropical Plant Biology*.
849 7:71–83.

850 Bottino MC, Rosario S, Grativol C, Thiebaut F, Rojas CA, Farrineli L, Hemerly AS, Ferreira
851 PCG. 2013. High-throughput sequencing of small RNA transcriptome reveals salt stress
852 regulated microRNAs in sugarcane. *PloS One*, 8:p.e59423.

853 Brühl R. 2015. A Mac OS X application for 2D-plots from ASCII data.
854 <http://rbruehl.macbay.de/>. Accessed 22 February 2019.

855 Bullerwell CE, Gray MW. 2004. Evolution of the mitochondrial genome: protist connections to
856 animals, fungi and plants. *Current Opinion in Microbiology*. 7:528–534.

857 Calsa Jr T, Carraro DM, Benatti MR, Barbosa AC, Kitajima JP, Carrer H. 2004. Structural
858 features and transcript-editing analysis of sugarcane (*Saccharum officinarum* L.) chloroplast
859 genome. *Current Genetics*. 46:366–373.

860 Carlsson J, Leino M, Sohlberg J, Sundstrom JF, Glimelius, K. 2008. Mitochondrial regulation of
861 flower development. *Mitochondrion* 8: 74–86.

862 Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and
863 additional sequence information. In: German Conference on Bioinformatics 99:45–56.

864 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2008.
865 BLAST+: architecture and applications. *BMC Bioinformatics*. 10:421.

866 Chang S, Yang T, Du T, Huang Y, Chen J, Yan J, He J, Guan R. 2011. Mitochondrial genome
867 sequencing helps show the evolutionary mechanism of mitochondrial genome formation in
868 Brassica. *BMC Genomics*. 12:497.

869 Chen Z, Zhao N, Li S, Grover CE, Nie H, Wendel JF, Hua J. 2017. Plant Mitochondrial Genome
870 Evolution and Cytoplasmic Male Sterility, *Critical Reviews in Plant Sciences*, 36:55-69. DOI:
871 10.1080/07352689.2017.1327762

872 Christensen AC. 2013. Plant mitochondrial genome evolution can be explained by DNA repair
873 mechanisms. *Genome Biology and Evolution*, 5:1079–1086.

874 Clayton WD, Vorontsova MS, Harman KT, Williamson H. (2006 onwards). GrassBase — The
875 Online World Grass Flora. <http://www.kew.org/data/grasses-db.html>. [accessed 10 May 2018;
876 15:30 GMT]

877 Clifton SW, Minx P, Fauron CMR, Gibson M Allen JO, Sun H, Thompson M, Barbazuk WB,
878 Kanuganti S, Tayloe C, Meyer L. 2004. Sequence and comparative analysis of the maize NB
879 mitochondrial genome. *Plant Physiology*, 136:3486-3503.

880 Conant GC, Wolfe KH. 2008. GenomeVx: simple web-based creation of editable circular
881 chromosome maps. *Bioinformatics*, 24:861–862

882 Daniels J, Roach BT. 1987. Taxonomy and evolution. In: D. J. Heinz Ed. *Sugarcane*
883 *Improvement Through Breeding*. Elsevier, The Hague. pp. 7–84.

884 Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved
885 genomic sequence with rearrangements. *Genome Research*. 14:1394–1403.

886 Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics
887 and parallel computing. *Nature Methods*, 9:772.

888 de Haas JM, Hille J, Kors F, van der Meer B, Kool AJ, Folkerts O, Nijkamp HJ. 1991. Two
889 potential *Petunia hybrida* mitochondrial DNA replication origins show structural and in vitro
890 functional homology with the animal mitochondrial DNA heavy and light strand replication
891 origins. *Current Genetics*. 20:503–513.

892 Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating
893 with confidence. *PLoS Biol*. 4:e88.

894 Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti
895 and the BEAST 1.7. *Molecular Biology and Evolution*, 29:1969–1973.

896 Fujii S, Kazama T, Ito Y, Kojima S, Toriyama K. 2014. A candidate factor that interacts with
897 RF2, a restorer of fertility of Lead rice-type cytoplasmic male sterility in rice. *Rice*, 7:21.

898 Gaborieau L, Brown GG, Mireau H. 2016. The propensity of pentatricopeptide repeat genes to
899 evolve into restorers of cytoplasmic male sterility. *Frontiers in Plant Science*, 7:1816.

900 Gagliardi D, Stepien PP, Temperley RJ, Lightowlers RN, Chrzanowska-Lightowlers ZM. 2004.
901 Messenger RNA stability in mitochondria: different means to an end. *Trends in Genetics*,
902 20:260-267.

903 Galtier N. 2011. The intriguing evolutionary dynamics of plant mitochondrial DNA. *BMC*
904 *Biology*. 9:61.

905 Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten
906 U, Putnam N, Rokhsar DS (2011) Phytozome: a comparative platform for green plant
907 genomics. *Nucleic Acids Research*. 40:D1178-D1186.

908 Grativol C, Regulski M, Bertalan M, McCombie WR, Silva FR, Zerlotini Neto A, Vicentini R,
909 Farinelli L, Hemerly AS, Martienssen RA, Ferreira PCG. 2014. Sugarcane genome sequencing
910 by methylation filtration provides tools for genomic research in the genus *Saccharum*. *The Plant*
911 *Journal*, 79:162-172.

912 Gualberto JM, Milesina D, Wallet C, Niazi AK, Weber-Lotfi F, Dietrich A. 2014. The plant
913 mitochondrial genome: dynamics and maintenance. *Biochimie*, 100:107-120.

914 Gualberto JM, Newton KJ. 2017. Plant mitochondrial genomes: dynamics and mechanisms of
915 mutation. *Annual Reviews of Plant Biology*. 68:17.1–17.28.

916 Guo W, Grewe F, Fan W, Young GJ, Knoop V, Palmer JD, Mower JP. 2016. Ginkgo and
 917 Welwitschia mitogenomes reveal extreme contrasts in gymnosperm mitochondrial evolution.
 918 *Molecular Biology and Evolution*. 33:1448-1460.

919 Hamani K, Giege P. 2014. RNA metabolism in plant mitochondria. *Trends in Plant Science*.
 920 19:380-389.

921 Hoang NV, Furtado A, McQualter RB, Henry RJ. 2015. Next generation sequencing of total
 922 DNA from sugarcane provides no evidence for chloroplast heteroplasmy. *New Negatives in*
 923 *Plant Science*, 1:33-45.

924 Huang WC, Yu CC, Hu J, Wang LL, Dan ZW, Zhou W, He CL, Zeng YF, Yao GX, Qi JZ,
 925 Zhang ZH, Zhu RS, Chen XF, Zhu YG. 2015. Pentatricopeptide-repeat family protein RF6
 926 functions with hexokinase 6 to rescue rice cytoplasmic male sterility. *Proceedings of the National*
 927 *Academy of Sciences U.S.A.* 112:14984–14989.

928 Huang X, Madan, A. CAP3: A DNA Sequence Assembly Program. 1999. *Genome Research*.
 929 9:868–877.

930 Hunt M, De Silva N, Otto TD, Parkhill J, Keane JA, Harris SR. 2015. Circlator: automated
 931 circularization of genome assemblies using long sequencing reads. *Genome Biology*. 16:294.

932 Igarashi K, Kazama T, Motomura K, Toriyama K. 2012. Whole genomic sequencing of RT98
 933 mitochondria derived from *Oryza rufipogon* and northern blot analysis to uncover a cytoplasmic
 934 male sterility-associated gene. *Plant and cell physiology*, 54:237-243.

935 Irvine JE. 1999. *Saccharum* species as horticultural classes. *Theoretical and Applied Genetics*.

936 98:186–194.

937 Itabashi E, Iwata N, Fujii S, Kazama T, Toriyama K, 2011. The fertility restorer gene, Rf2, for
 938 Lead Rice-type cytoplasmic male sterility of rice encodes a mitochondrial glycine-rich protein.
 939 *The Plant Journal*, 65:359-367.

940 Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre² web portal for
 941 protein modeling, prediction and analysis. *Nature Protocols*, 10:845.

942 Kersey PJ, Allen JE, Allot A, Barba M, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M,
 943 Davis P, Grabmueller C, Kumar N. 2017. Ensembl Genomes 2018: an integrated omics
 944 infrastructure for non-vertebrate species. *Nucleic Acids Research* 46:D802-D808.

945 Kim D, Langmead B, Salzberg, SL. 2015. HISAT: a fast spliced aligner with low memory
 946 requirements. *Nature Methods*. 12:357–360.

947 Kohany O, Gentles AJ, Hankus L and Jurka J. 2006. Annotation, submission and screening of
 948 repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, 7:474.

949 Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH and Phillippy AM. 2017. Canu: scalable
 950 and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome*
 951 *Research*. 27:722-736

952 Krishnan NM, Rao BJ. 2009. A comparative approach to elucidate chloroplast genome
 953 replication. *BMC Genomics*. 10:237. doi:10.1186/1471-2164-10-237.

954 Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. 2001. REPuter:
955 the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*.
956 29:4633–4642.

957 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform.
958 *Bioinformatics*. 25:1754-1760.

959 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.
960 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25:2078-2079.

961 Lima MS, Smith DR. 2017. Pervasive Transcription of Mitochondrial, Plastid, and Nucleomorph
962 Genomes across Diverse Plastid-Bearing Species. *Genome Biology and Evolution*. 9:2650–2657.
963 doi:10.1093/gbe/evx207

964 Lister DL, Bateman JM, Purton S, Howe CJ. 2003. DNA transfer from chloroplast to nucleus is
965 much rarer in *Chlamydomonas* than in tobacco. *Gene*. 316:33–38.

966 Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T. 2009. Rapid and accurate large-scale
967 coestimation of sequence alignments and phylogenetic trees. *Science*. 324:1561–1564.

968 Lloyd Evans D, Joshi SV. 2016. Complete chloroplast genomes of *Saccharum spontaneum*,
969 *Saccharum officinarum* and *Miscanthus floridulus* (Panicoideae: Andropogoneae) reveal the
970 plastid view on sugarcane origins. *Systematics and Biodiversity*. 14:548–571.

971 Lloyd Evans D, Joshi SV. 2017. Herbicide targets and detoxification proteins in sugarcane: from
972 gene assembly to structure modelling. *Genome*. 60:601-617.

973 Lloyd Evans D, Joshi SV, Wang J. 2019. Whole chloroplast genome and gene locus
974 phylogenies reveal the taxonomic placement and relationship of *Tripidium* (Panicoideae:
975 Andropogoneae) to sugarcane. *BMC Evolutionary Biology* 19:33.

976 Lloyd Evans D. 2019. A novel protein-based approach to transcriptome assembly in orphan
977 species. Validation on herbicide targets and low copy number genes in Gymnosperms, Juncaceae
978 and Pteridophyta. *Plant Molecular Biology*. In press.

979 Mackenzie S, McIntosh L. 1999. Higher plant mitochondria. *The Plant Cell*, 11:571–585.

980 Manchekar M, Scissum-Gunn K, Song D, Khazi F, McLean SL, Nielsen BL. 2006. DNA
981 recombination activity in soybean mitochondria. *Journal of Molecular Biology*. 356:288–299.

982 Mattiello L, Riaño-Pachón DM, Martins MCM, da Cruz LP, Bassi D, Marchiori PER, Ribeiro
983 RV, Labate MTV, Labate CA, Menossi M. 2015. Physiological and transcriptional analyses of
984 developmental stages along sugarcane leaf. *BMC Plant Biology*, 15:300.

985 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,
986 Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a
987 MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*.
988 20:1297–1303.

989 Mennes CB, Smets EF, Moses SN, Merckx VS. 2013. New insights in the long-debated
990 evolutionary history of *Triuridaceae* (Pandanales). *Molecular Phylogenetics and Evolution*.
991 69:994–1004.

992 Mower JP, Case AL, Floro ER, Willis JH. 2012. Evidence against equimolarity of large repeat
993 arrangements and a predominant master circle structure of the mitochondrial genome from a

994 monkeyflower (*Mimulus guttatus*) lineage with cryptic CMS. *Genome Biology and Evolution*.
 995 4:670–686.

996 Nylander JA, Wilgenbusch JC, Warren DL, Swofford DL. 2008. AWTY (are we there yet?): a
 997 system for graphical exploration of MCMC convergence in Bayesian phylogenetics.
 998 *Bioinformatics*. 24:581–583.

999 Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi
 1000 M, Chang Z, Aota SI. 1986. Chloroplast gene organization deduced from complete sequence of
 1001 liverwort *Marchantia polymorpha* chloroplast DNA. *Nature*. 322:572–574.

1002 Otto TD, Dillon GP, Degraeve WS, Berriman M. (2011). RATT: Rapid Annotation Transfer
 1003 Tool. *Nucleic Acids Research*. doi: 10.1093/nar/gkq1268

1004 Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004.
 1005 UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of*
 1006 *Computational Chemistry*, 25:1605-1612.

1007 Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005.
 1008 InterProScan: protein domains identifier. *Nucleic Acids Research*, 33:W116-W120.

1009 Rambaut A, Suchard MA, Xie D, Drummond AJ. 2013. Tracer, Version 1.5. Available at:
 1010 tree.bio.ed.ac.uk/software/tracer/.

1011 Reddy, B.V., Ramesh, S., Kumar, A.A., Wani, S.P., Ortiz, R., Ceballos, H. and Sreedevi T.K.
 1012 2008. Bio-fuel crops research for energy security and rural development in developing countries.
 1013 *Bioenergy Res*. 1:248–258.

1014 Riaño-Pachón DM, Mattiello L. 2017. Draft genome sequencing of the sugarcane hybrid SP80-
1015 3280. *F1000Research*, 6: 11859.2. DOI:10.12688/f1000research.11859.2

1016 Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open
1017 Software Suite. *Trends in Genetics*. 16:276–277.

1018 Ronquist F, Huelsenbeck JP 2003. MrBayes 3: Bayesian phylogenetic inference under mixed
1019 models. *Bioinformatics*. 19:1572–1574.

1020 Sanchez-Puerta MV, Garcia LE, Wohlfeiler J, Ceriotti LF. 2017. Unparalleled replacement of
1021 native mitochondrial genes by foreign homologs in a holoparasitic plant. *New Phytologist*. 214:
1022 376–387.

1023 Savojardo, C., Martelli, P.L., Fariselli, P. and Casadio, R., 2014. TPpred2: improving the
1024 prediction of mitochondrial targeting peptide cleavage sites by exploiting sequence motifs.
1025 *Bioinformatics*, 30(20), pp.2973-2974.

1026 Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for
1027 the detection of tRNAs and snoRNAs. 2005. *Nucleic Acids Research*. 33:W686–689 .

1028 Schmitz-Linneweber C, Small I. 2008. Pentatricopeptide repeat proteins: a socket set for
1029 organelle gene expression. *Trends in Plant Science*, 13:663-670.

1030 Shearman JR, Sonthirod C, Naktang C, Pootakham W, Yoocha T, Sangsrakru D, Jomchai N,
1031 Tragoonrung S, Tangphatsornruang S. 2016. The two chromosomes of the mitochondrial
1032 genome of a sugarcane cultivar: assembly and recombination analysis using long PacBio reads.
1033 *Scientific Reports*. 6:31533.

1034 Shi C, Wang S, Xia EH, Jiang JJ, Zeng FC, Gao LZ. 2016. Full transcription of the chloroplast
1035 genome in photosynthetic eukaryotes. *Scientific Reports*. 6:30135.

1036 Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence
1037 comparison. *BMC Bioinformatics*. 6:31; doi: 10.1186/1471-2105-6-31

1038 Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR. 2012.
1039 Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with
1040 exceptionally high mutation rates. *PLoS Biology*. 10:e1001241.

1041 Sobhakumari VP. 2013. New determinations of somatic chromosome number in cultivated
1042 and wild species of *Saccharum*. *Caryologia: International Journal of Cytology,*
1043 *Cytosystematics and Cytogenetics*. 66:268–274.

1044 Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with
1045 thousands of taxa and mixed models. *Bioinformatics*. 22:2688–2690.

1046 Subramanyam KN, Andal R. 1984. Male Sterility in Sugarcane. *Current Science*. 53:42-43.

1047 Sukumaran J, Holder MT. 2010. DendroPy: A Python library for phylogenetic
1048 computing. *Bioinformatics*. 26:1569-1571.

1049 Suzuki H, Yu J, Ness S, O’Connell M, Zhang J. 2013. RNA editing events in mitochondrial
1050 genes by ultra-deep sequencing methods: a comparison of cytoplasmic male sterile, fertile and
1051 restored genotypes in cotton. *Mol. Genet. Genomics*. 288: 445–457.

1052 Taylor TC, Andersson I. 1997. The structure of the complex between rubisco and its natural
1053 substrate ribulose 1, 5-bisphosphate. *Journal of Molecular Biology*, 265:432-444.

1054 Thorvaldsson H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-
1055 performance genomics data visualization and exploration. *Briefings in Bioinformatics*. 14:178–
1056 192.

1057 Toriyama K, Kazama T, Motomura K, Igarishi K. 2013. Rice RT-type Cytoplasmic Male
1058 Sterility Causal Gene and Use Thereof. International Patent Application WO2014027502.
1059 [https://patents.google.com/patent/WO2014027502A1/en?q=Rice+RT-](https://patents.google.com/patent/WO2014027502A1/en?q=Rice+RT-type+Cytoplasmic+Male+Sterility+Causal+Gene+and+Use+Thereof.+International+Patent+Application+WO2014027502)
1060 [type+Cytoplasmic+Male+Sterility+Causal+Gene+and+Use+Thereof.+International+Patent](https://patents.google.com/patent/WO2014027502A1/en?q=Rice+RT-type+Cytoplasmic+Male+Sterility+Causal+Gene+and+Use+Thereof.+International+Patent+Application+WO2014027502)
1061 [+Application+WO2014027502](https://patents.google.com/patent/WO2014027502A1/en?q=Rice+RT-type+Cytoplasmic+Male+Sterility+Causal+Gene+and+Use+Thereof.+International+Patent+Application+WO2014027502).

1062 Tsujimura M, Kaneko T, Sakamoto T, Kimura S, Shigyo M, Yamagishi M, Toru Terachi T. 2018.
1063 Multichromosomal structure of the onion mitochondrial genome and a transcript analysis. 2018.
1064 doi: 10.1016/j.mito.2018.05.001

1065 Turnel M, Otis C, Lemieux C. 2003. The mitochondrial genome of *Chara vulgaris*: insights into
1066 the mitochondrial DNA architecture of the last common ancestor of green algae and land plants.
1067 *The Plant Cell*. 15:1888–1903.

1068 Vargas L, Santa Brígida AB, Mota Filho JP, de Carvalho TG, Rojas CA, Vaneechoutte D, Van
1069 Bel M, Farrinelli L, Ferreira PC, Vandepoele K, Hemerly AS. 2014. Drought tolerance conferred
1070 to sugarcane by association with *Gluconacetobacter diazotrophicus*: a transcriptomic view of
1071 hormone pathways. *PLoS One*, 9:p.e114744.

1072 Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara
1073 GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome*
1074 *Research*. 19:327-335.

- 1075 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q,
1076 Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial
1077 variant detection and genome assembly improvement. *PloS One*. 9:p.e112963.

- 1078 Zhang, N., Rao, R.S.P., Salvato, F., Havelund, J.F., Møller, I.M., Thelen, J.J. and Xu, D., 2018.
1079 MU-LOC: a machine-learning method for predicting mitochondrially localized proteins in
1080 plants. *Frontiers in plant science*, 9.


- 1081 Zhang T, Hu S, Zhang G, Pan L, Zhang X, Al-Mssallem IS, Yu J. 2012. The organelle genomes
1082 of Hassawi rice (*Oryza sativa* L.) and its hybrid in Saudi Arabia: genome variation,
1083 rearrangement, and origins. *PloS One*, 7:p.e42041.

Figure Legends

Figure 1. Circular images of the *Saccharum* hybrid SP80-3280 mitochondrial genome.

Circular diagrams of the mitochondrial chromosomes of sugarcane hybrid cultivar SP80-3280. Bars on the outer circle represent genes (with forward strand genes on the outer track and reverse strand genes on the inner track). All genes are labelled and the large direct repeat (DR) and inverted repeats (IR) are shown and labelled on the centre track of chromosome 1. The inner, grey, circle represents GC content. Images were drawn with GenomeVX (Conant and Wolfe, 2008)

Figure 2. Phylogram and Chronogram generated from sugarcane mitochondrial chromosome 2 data.

A phylogram (left) was generated from mitochondrial chromosome 2 data for sugarcane and reads mapped to chromosome 2 for other species. The phylogram was generated with RAxML, and numbers above nodes represent maximum likelihood bootstrap support, whilst numbers below nodes represent Bayesian inference support. The scale bar at the bottom represents numbers of substitutions per site. The // mark represents long branches that have been reduced by 50%. The image ight, gives a chronogram generated with BEAST for the mitochondrial data. The scale axis (bottom) gives numbers in millions of years before present. The numbers at nodes represent the age of the node (in millions of years before present). Node bars represent 95% highest probability densities (HPD) on the age of the node.

1105

1106 **Figure 3.** The Spliceosome of the Sugarcane Mitochondrion

1107 Image of the complete spliceosome of the sugarcane mitochondrion drawn with IGV.

1108 Chromosome 1 and chromosome 2 are concatenated together in this view but the extents

1109 of MT1 and MT2 are marked. Both strands are shown and spliceosomal events occur when

1110 the red and blue lines touch the line dividing the forward and reverse mapped reads. Splice

1111 sites typically seem to cluster in hotspots where there is considerable mapping depth.

1112 Though long range splice events predominate short-range splice events can still be seen

1113 (narrow humps in the background). The most common splice sites (boxed) are between the

1114 start of chromosome 1 and the start of chromosome 2. The denser the colour map the more

1115 splice sites span that region.

1116

1117

1118 **Figure 4.** PolyA Tailed Mitochondrial RNAs in Sorghum

1119 An image generated from IGV showing the mapping of polyA baited transcript reads to the

1120 Sorghum bicolor cv BTx623 mitochondrial genome. Regions of contiguous high mapping depth

1121 are boxed and numbered. A full analysis of the mapped regions, including the genes/features

1122 contained therein are available in Supplemental Table S4.

1123

1124 **Figure 5.** Mapping transcriptomic data to the sugarcane SP80-3280 mitochondrion and

1125 chloroplast.

1126 Image showing the results of mapping transcriptomic reads to the sugarcane SP80-3280
1127 mitochondrial and chloroplast genomes. A: SP80-3280 mitochondrial chromosome 1. B: SP80-
1128 3280 mitochondrial chromosome 2. C: SP80-3280 chloroplast genome. The y-axis represents
1129 \log_{10} counts for transcript coverage at each base position within the genome. The x-axis
1130 represents base position within the genome.

1131

1132

1133 **Figure 6.** Structural comparisons of the sugarcane chloroplast and mitochondrial version of
1134 rbcL.

1135 Superimposition and structural comparisons of the sugarcane chloroplast (mauve) and
1136 mitochondrial (green) version of the rbcL (RuBisCO large subunit). As can be seen, the
1137 structures are virtually identical and apart from truncations in the disordered amino (N)
1138 and carboxy (C) termini of the mitochondrial protein the only meaningful difference is the
1139 prediction of a helix centred on R86 in the chloroplast molecule and the prediction of a
1140 corresponding loop centred on Arg79 in the mitochondrial protein (shown with an arrow).
1141 However, as the sequences in the two regions are identical, this difference is almost
1142 certainly not meaningful. Otherwise, the structures are identical and active site amino
1143 acids are conserved; a strong indication that the sugarcane mitochondrial version of rbcL
1144 could be functional.

1145

1146 **Figure 7.** Domain and protein feature mappings of the sugarcane mitochondrial CMS factor and

1147 three putative genomic restoration factors.

1148 Images represent: A) the sugarcane mitochondrial CMS factor, showing the extent of the first;
 1149 transmembrane, helix as predicted by TMHMM and the Transmembrane region as predicted by
 1150 PHOBIUS as implemented in InterProScan (Quevillon et al., 2005); B) ShRf11, a potential
 1151 restorer of function 1 like transcript, showing the mitochondrial transit peptide and all the PPR
 1152 (pentatricopeptide) repeats within the protein; C) the sugarcane orthologue of rice and sorghum
 1153 DSK2 protein, a restorer of function gene with an ubiquitin superfamily domain at the N-
 1154 terminus and an UBA-like domain responsible for polypeptide substrate binding at the C-
 1155 terminus and A) the ShRf21 (restorer of function 2 like) protein, which has no recognised
 1156 domains, but which does contain a conserved glycine-rich region.

1157

Table Legends

Table 1. Primers used to amplify transcripts from the SP80-3280 and N22 sugarcane cDNA libraries.

A list of primers used to amplify potential restorer of function transcripts in both the SP80-3280 and N22 sugarcane cultivars. This table gives the gene names and types for the three potential CMS restorer of function transcripts identified in sugarcane. Also given are the forward and reverse primers used to amplify the transcripts, the length of the amplicons obtained and the melting temperatures (T_m) for the primers.

Table 2. Comparisons of base-level differences in the mitochondria and chloroplasts of sugarcane cultivars to the SP80-3280 reference assemblies presented in this paper.

Analysis of base-by base comparisons of several sugarcane mitochondrial and chloroplast assemblies from different cultivars to the reference SP80-3280 assemblies presented in this paper. Mitochondrial data is given at the top and chloroplast data at the bottom. Columns represent: cultivar; total length of plastome; total number of substitutions; total number of insertions; total number of deletions. For mitochondria, positions of large direct and inverted repeats and the total number of small repeats are given. Numbers in brackets give substitutions corrected for transcript post-processing. The label ‘gb’ means that the sequence is one downloaded from GenBank.

Supplemental Materials

Supplemental Table S1

Summary of GC content in chloroplast, mitochondrial and nuclear genomes of a representative sample of Saccharinae. Table headers represent: species, species name; voucher or accession, species voucher, cultivar name or accession; GenBank accession, GenBank identifier of sequence (if available); reference, published reference for the sequence and GC Content (%), percentage GC content in the genome.

Supplemental Table S2

Transposable elements in the sugarcane SP80-3280 mitochondrial genome. Transposable elements were derived from the Poaceae dataset of Censor. Column headings: Left and right, position of the transposable element in the mitochondrial genome (From/To indicates start/end of positions of the transposable elements). Orientation: + forward strand; -, complementary. 'Sim' indicates value of similarity between 2 aligned fragments; 'Pos' is the ratio of positives to alignment length; 'Mm:Ts' is a ratio of mismatches to transitions in the nucleotide alignment. 'Score', alignment score obtained from blast.

Supplemental Table S3

A complete listing of splice sites in the sugarcane mitochondrial genome. A listing of all splice sites with occurrence > 10 in the sugarcane mitochondrial genome. The table lists the splice

1200 sites positions, upstream and downstream genes and the notes give any coding sequences that
1201 the splice sites lie within.

1202

1203 **Supplemental Table S4**

1204 Table listing polyA baited read mapping to the Sorghum bicolor BTx623 mitochondrial genome.
1205 Table lists the start and end of all contiguous polyA read mapping sites with depth >500 reads
1206 along with any genes or notable featured within the mapping region.

1207

1208 **Supplemental Document S1**

1209 Mapping of RNA-seq reads to the multi-chromosomal mitogenomes of three species. Document
1210 gives column graphs of site by site mappings of RNA-seq reads to each chromosome in the
1211 multi-chromosome mitogenomes of *Silene vulgaris*, *Cucumis sativus* and *Allium cepa* showing
1212 almost complete coverage of each chromosome. The document lists all the SRA datasets
1213 employed in each of the mappings.

Figure 1(on next page)

Circular images of the *Saccharum* hybrid SP80-3280 mitochondrial genome.

Circular diagrams of the mitochondrial chromosomes of sugarcane hybrid cultivar SP80-3280. Bars on the outer circle represent genes (with forward strand genes on the outer track and reverse strand genes on the inner track). All genes are labelled and the large direct repeat (DR) and inverted repeats (IR) are shown and labelled on the centre track of chromosome 1. The inner, grey, circle represents GC content. Images were drawn with GenomeVX (Conant and Wolfe, 2008).

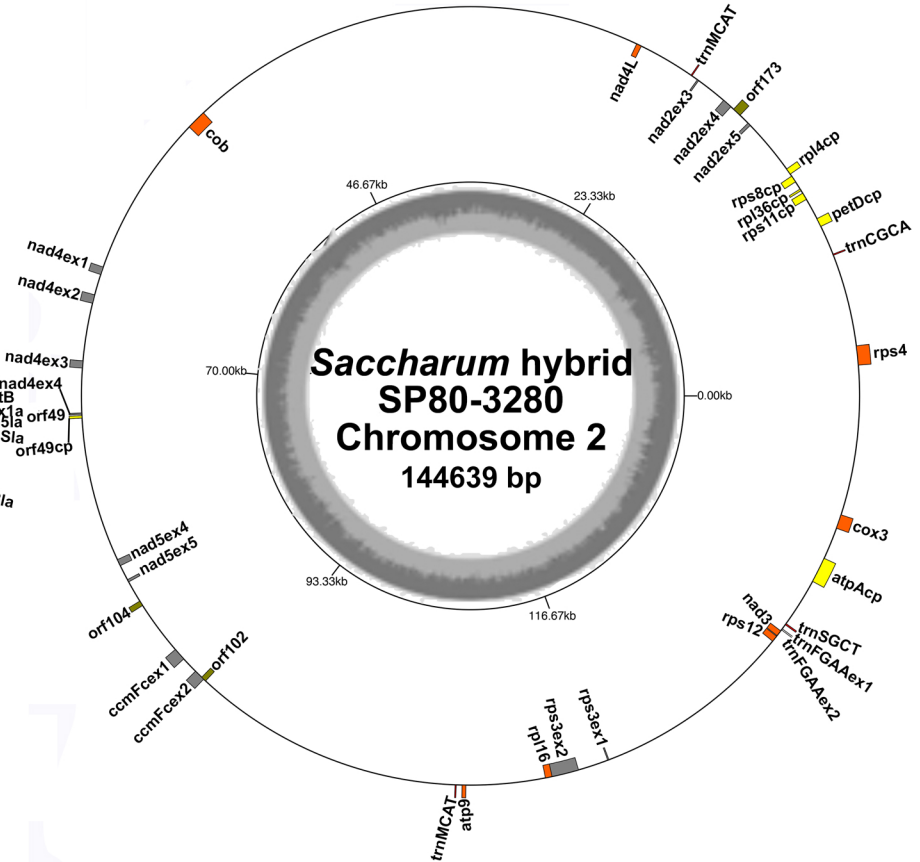
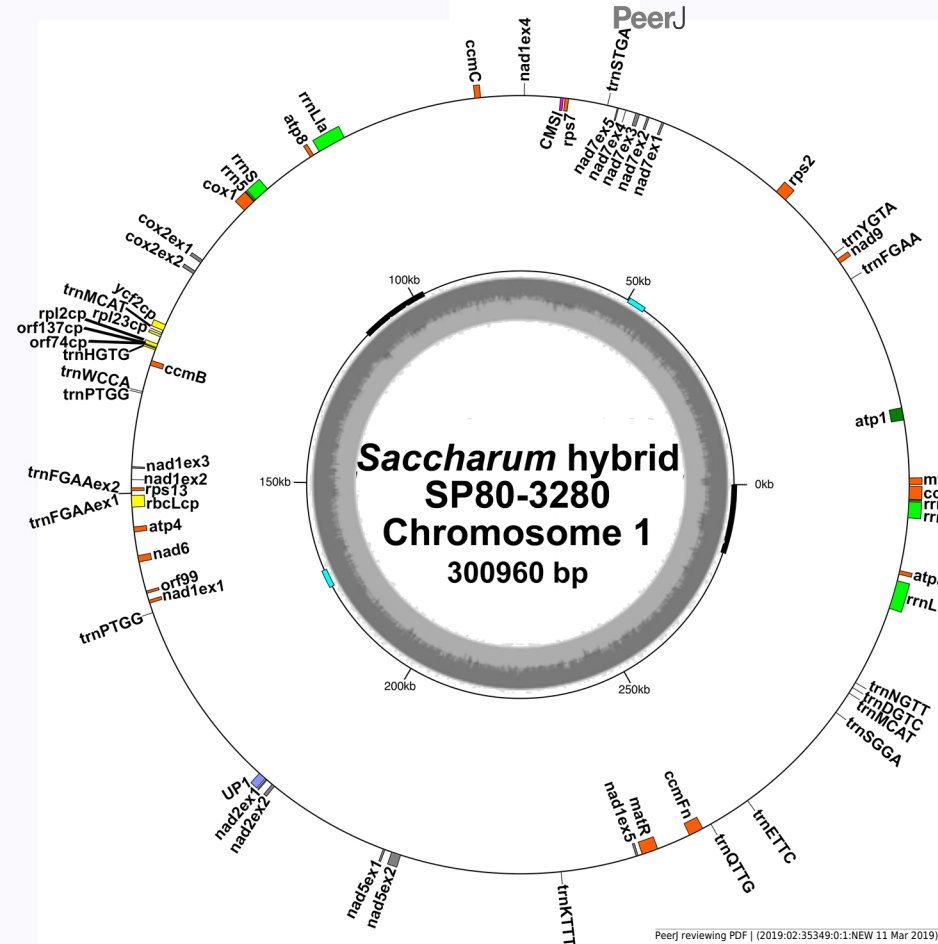


Figure 2 (on next page)

Phylogram and Chronogram generated from sugarcane mitochondrial chromosome 2 data.

A phylogram (left) was generated from mitochondrial chromosome 2 data for sugarcane and reads mapped to chromosome 2 for other species. The phylogram was generated with RAxML, and numbers above nodes represent maximum likelihood bootstrap support, whilst numbers below nodes represent Bayesian inference support. The scale bar at the bottom represents numbers of substitutions per site. The // mark represents long branches that have been reduced by 50%. The image, right, gives a chronogram generated with BEAST for the mitochondrial data. The scale axis (bottom) gives numbers in millions of years before present. The numbers at nodes represent the age of the node (in millions of years before present). Node bars represent 95% highest probability densities (HPD) on the age of the node.

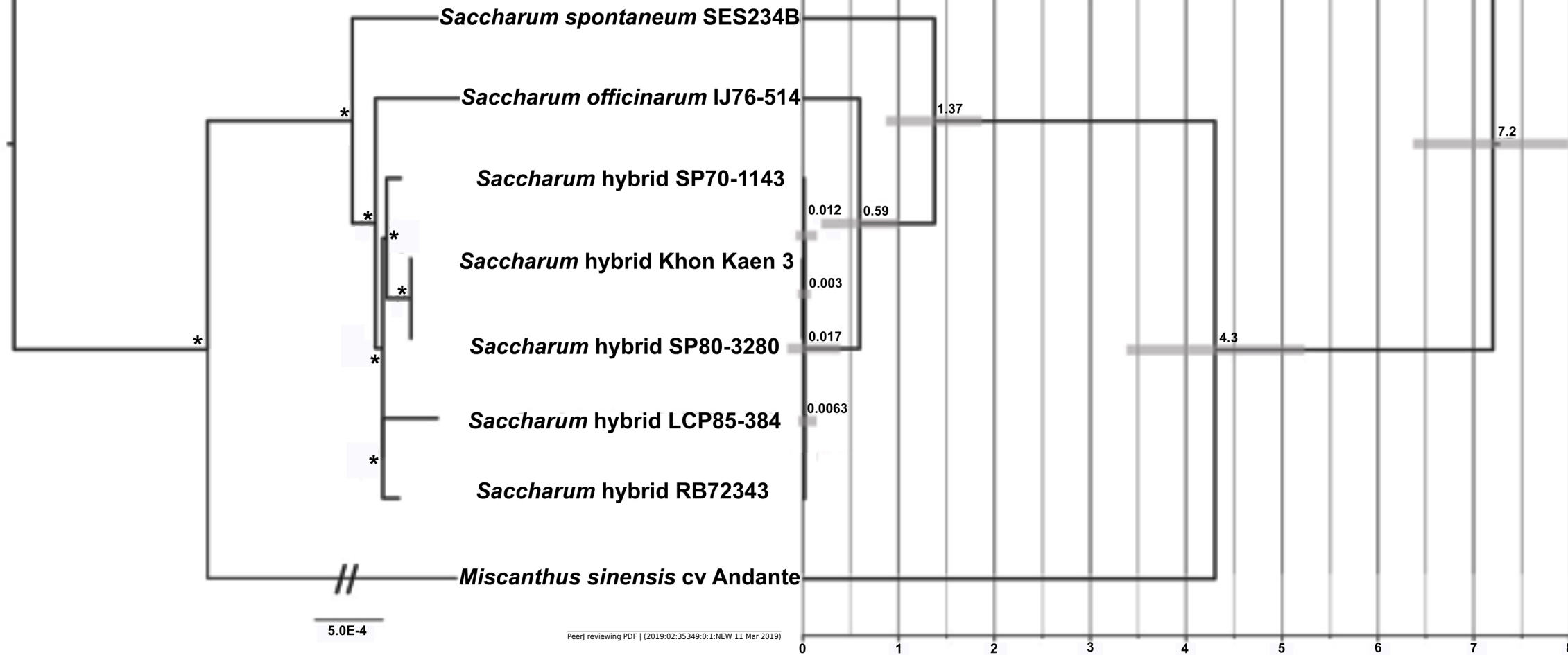


Figure 3(on next page)

The Spliceosome of the Sugarcane Mitochondrion

Image of the complete spliceosome of the sugarcane mitochondrion drawn with IGV.

Chromosome 1 and chromosome 2 are concatenated together in this view but the extents of MT1 and MT2 are marked. Both strands are shown and spliceosomal events occur when the red and blue lines touch the line dividing the forward and reverse mapped reads. Splice sites typically seem to cluster in hotspots where there is considerable mapping depth. Though long range splice events predominate short-range splice events can still be seen (narrow humps in the background). The most common splice sites (boxed) are between the start of chromosome 1 and the start of chromosome 2. The denser the colour map the more splice sites span that region

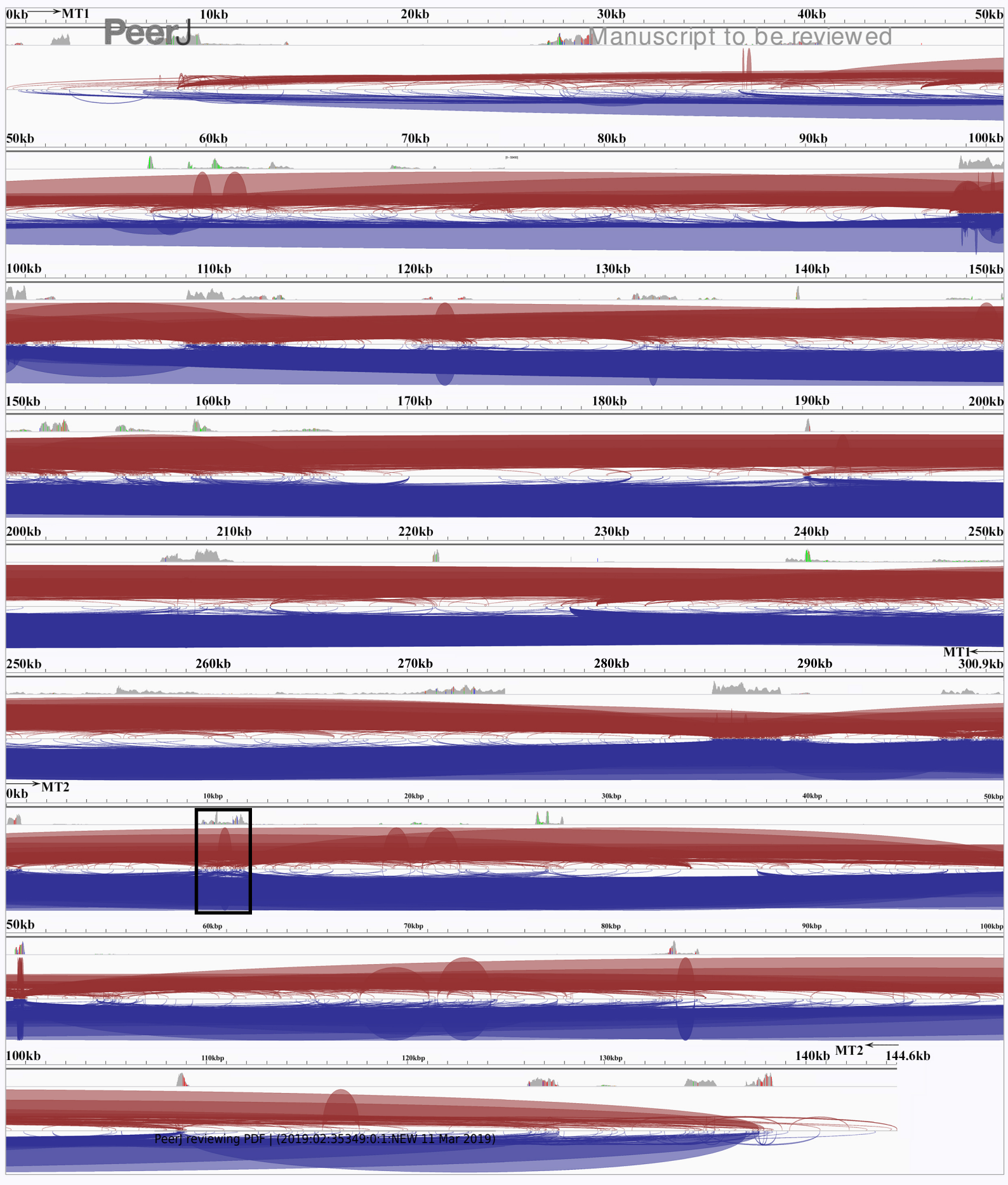


Figure 4(on next page)

PolyA-Tailed Mitochondrial RNAs in *Sorghum*

An image generated from IGV showing the mapping of polyA-baited transcript reads to the *Sorghum bicolor* cv BTx623 mitochondrial genome. Regions of contiguous high mapping depth are boxed and numbered. A full analysis of the mapped regions, including the genes/features contained therein is available in Supplemental Table S4.

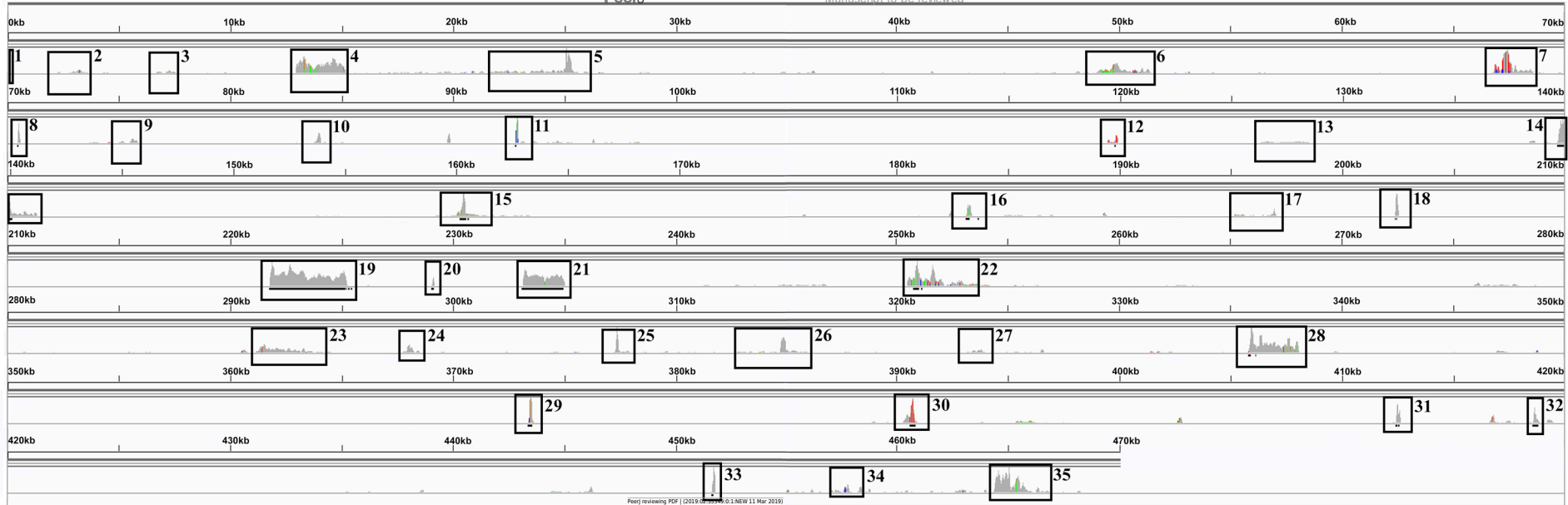


Figure 5(on next page)

Mapping transcriptomic data to the sugarcane SP80-3280 mitochondrion and chloroplast.

Image showing the results of mapping transcriptomic reads to the sugarcane SP80-3280 mitochondrial and chloroplast genomes. A: SP80-3280 mitochondrial chromosome 1. B: SP80-3280 mitochondrial chromosome 2. C: SP80-3280 chloroplast genome. The y-axis represents \log_{10} counts for transcript coverage at each base position within the genome. The x-axis represents base position within the genome.

A

PeerJ

Manuscript to be reviewed

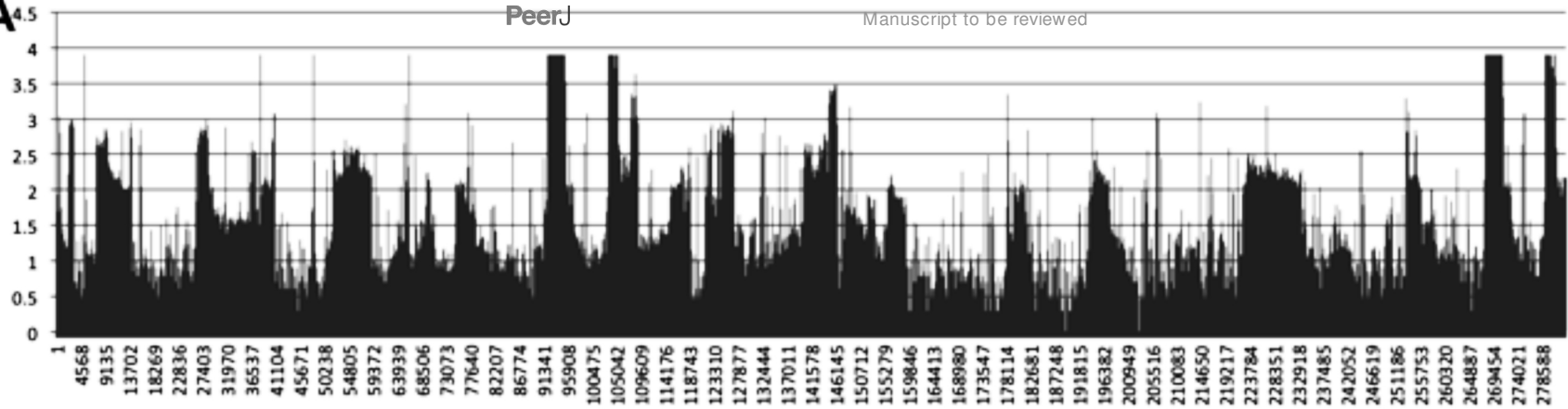
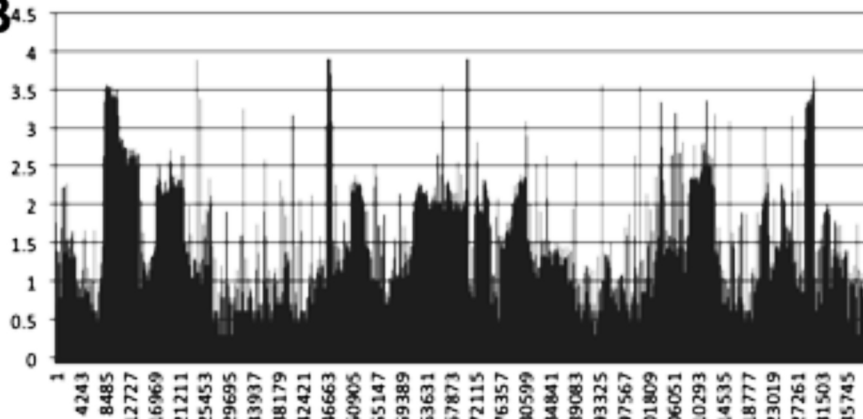
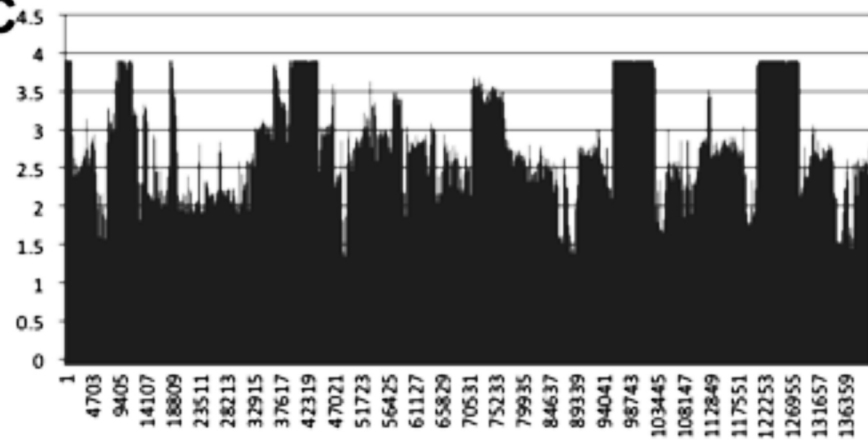
**B****C**

Figure 6(on next page)

Structural comparisons of the sugarcane chloroplast and mitochondrial version of rbcL

Superimposition and structural comparisons of the sugarcane chloroplast (mauve) and mitochondrial (green) version of the rbcL (RuBisCO large subunit protein). As can be seen, the structures are virtually identical and apart from truncations in the disordered amino (N) and carboxyl (C) termini of the mitochondrial protein the only meaningful difference is the prediction of a helix centred on R86 in the chloroplast molecule and the prediction of a corresponding loop centred on Arg79 in the mitochondrial protein (shown with an arrow). However, as the sequences in the two regions are identical, this difference is almost certainly artifactual and not meaningful. Otherwise, the structures are identical and active site along with substrate contact amino acids are conserved; a strong indication that the sugarcane mitochondrial version of rbcL could be functional.

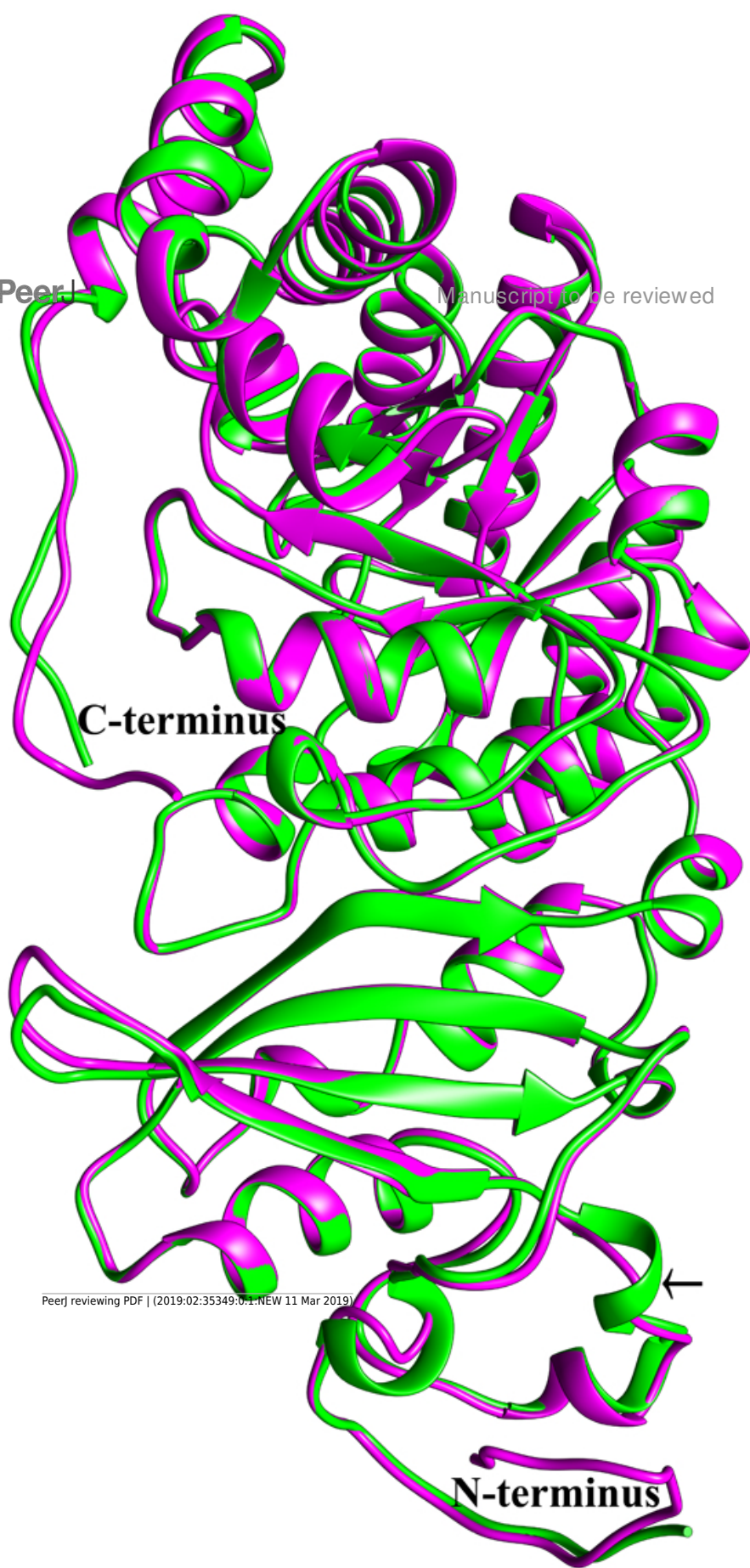


Figure 7 (on next page)

Domain and protein feature mappings of the sugarcane mitochondrial CMS factor and three putative genomic restoration factors.

Images represent: A) the sugarcane mitochondrial CMS factor, showing the extent of the first, transmembrane, helix as predicted by TMHMM and the Transmembrane region as predicted by PHOBIUS as implemented in InterProScan (Quevillon et al., 2005); B) ShRf1I, a potential restorer of function 1 like transcript, showing the mitochondrial transit peptide and all the PPR (pentatricopeptide) repeats within the protein; C) the sugarcane orthologue of rice and sorghum DSK2 protein, a restorer of function gene with an ubiquitin superfamily domain at the N-terminus and an UBA-like domain responsible for polypeptide substrate binding at the C-terminus; and D) the ShRf2I/GRP162 (restorer of function 2 like/glycine rich RNA-binding protein 3) protein, which has a conserved RRM (RNA recognition motif) domain, along with a conserved glycine-rich repeat region.

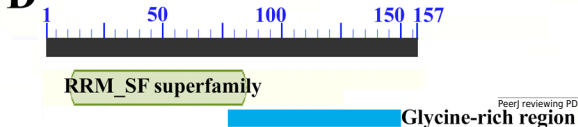
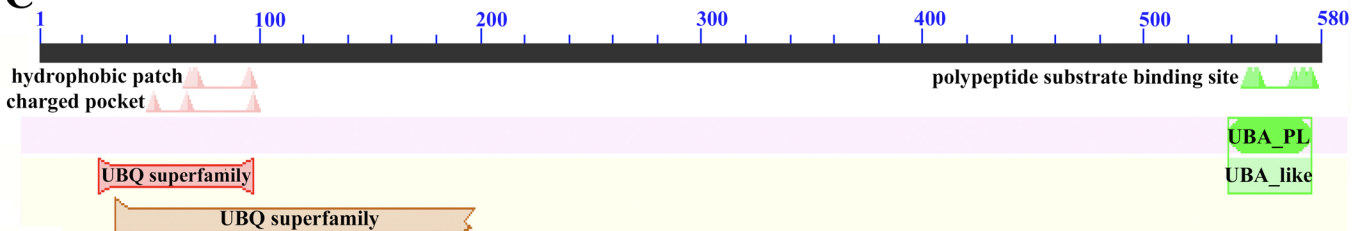
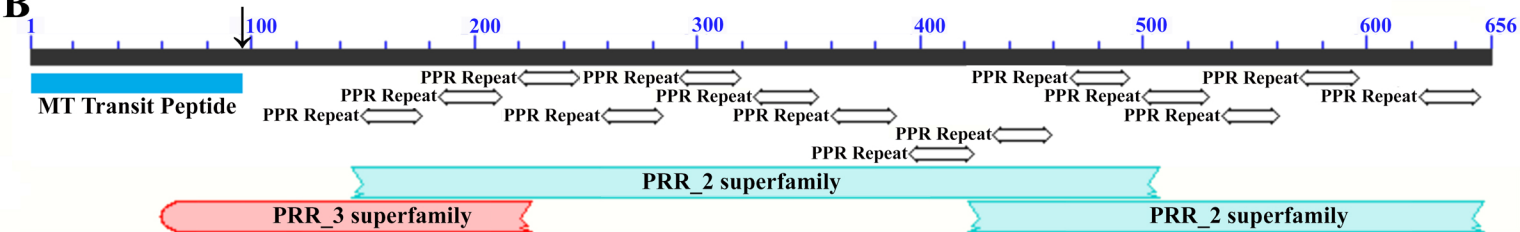


Figure 8(on next page)

Multiple sequence alignment of mitogenome derived rbcL against the chloroplast reference.

Multiple sequence alignment comparing the protein sequences of mitogenome derived rbcL from *Saccharum* species and sugarcane cultivars against the reference chloroplastic. rbcL of sugarcane cultivar SP80-3280. The rbcL sequence from *Saccharum* hybrid RB72454 was identical to that of LCP85-384 and is not shown in the alignment. The chloroplast-derived sequence is shown for reference at the top. All other sequences are given in phylogenetic order.

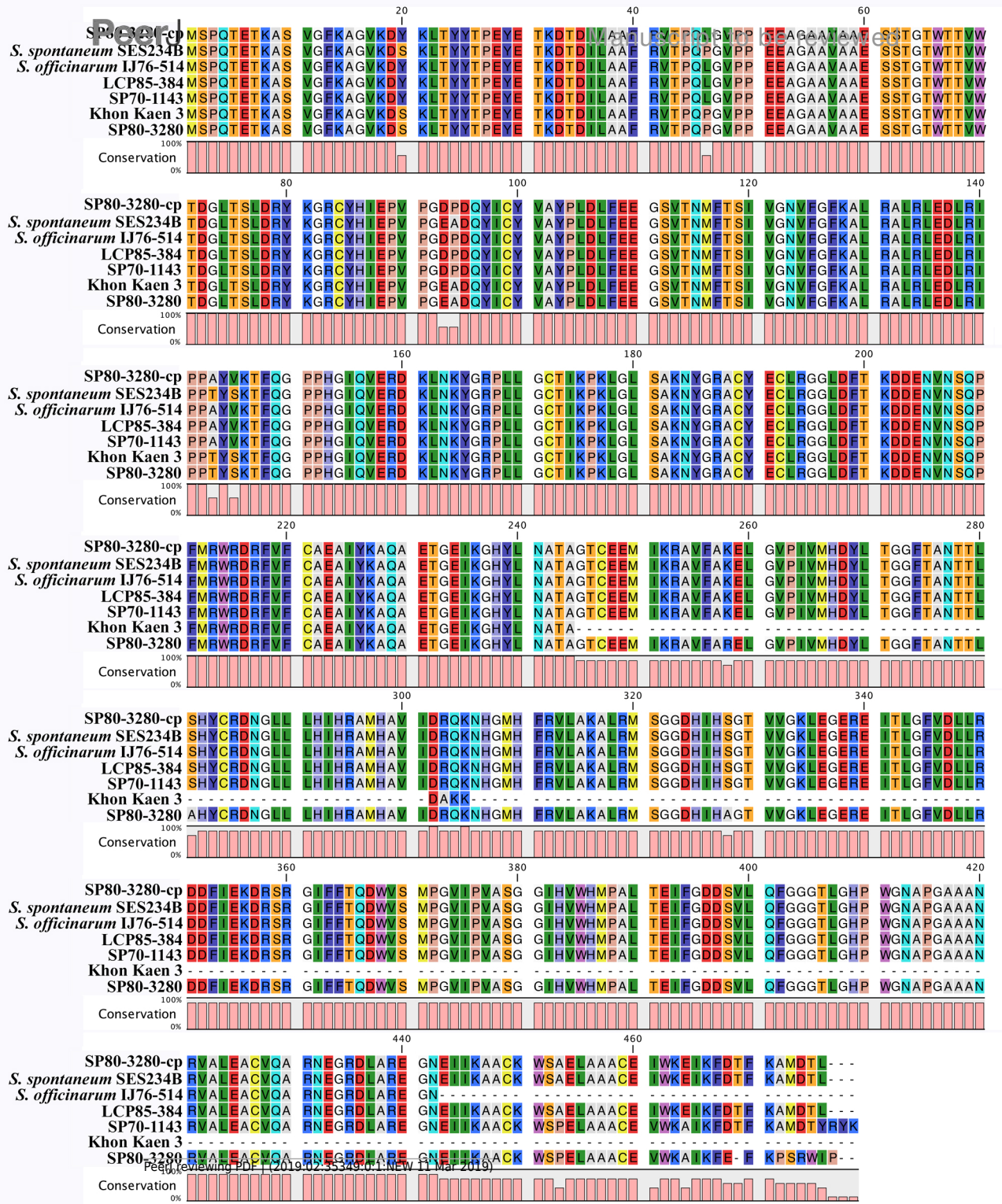


Table 1(on next page)

Primers used to amplify transcripts from the SP80-3280 and N22 sugarcane cDNA libraries.

A list of primers used to amplify potential restorer of function transcripts in both the SP80-3280 and N22 sugarcane cultivars. This table gives the gene names and types for the three potential CMS restorer of function transcripts identified in sugarcane. Also given are the forward and reverse primers used to amplify the transcripts, the length of the amplicons obtained and the melting temperatures (T_m) for the primers.

Gene	Left Primer	Right Primer	SP80 Amplicon Length	N22 Amplicon Length	Tm
ShRF1 PPR domain protein	GCGCGACCGAGCTGCATTTCC	TCCCCTTTTGGCCATCTGCAGC	2133	2136	72°C
ShDSK2 ubiquitin domain protein	GGAACGAATCCGGACCGTC	TTGAAACCACCGGTTGGATTAG	2313	2312	63°C
ShGRP162 (glycine-rich RNA-binding protein 3)	GTGCGCGTAGCGCAGCGGGG	TGGCAGCACCAAGAAGCACCTTTTTTT	1030	1030	72°C

1

Table 2 (on next page)

Comparisons of base-level differences in the mitochondria and chloroplasts of sugarcane cultivars to the SP80-3280 reference assemblies presented in this paper.

Analysis of base-by base comparisons of several sugarcane mitochondrial and chloroplast assemblies from different cultivars to the reference SP80-3280 assemblies presented in this paper. Mitochondrial data is given at the top and chloroplast data at the bottom. Columns represent: cultivar; total length of plastome; total number of substitutions; total number of insertions; total number of deletions. For mitochondria, positions of large direct and inverted repeats and the total number of small repeats are given. Numbers in brackets give substitutions corrected for transcript post-processing. The label 'gb' means that the sequence is one downloaded from GenBank.

Mitochondria	Length	substitutions	insertions	deletions	Repeats		small repeats (<360 bp)
					15k	4k	
SP80-3284 mt1	300960				9777-285530	45748-174194R	146
SP80-3284 mt2	144639						48
IJ76-514 mt1	300995	470	25	8	98560-289970	45945-174355R	134
IJ76-514 mt2	144926	261	32	8			121
RB72454 mt1	300828	79	3	7	97558-285312	45748-174074R	139
RB72454 mt2	144692	67	8	1			52
LCP85-384 mt1	300775	126	1	3	97691-285426	46049-173891R	147
LCP85-384 mt2	144679	105	7	0			48
Khon Kaen 3 mt1 (gb)	300784	40	5	11	97558-288181	45748-174045R	147
Khon Kaen 3 mt2 (gb)	144648	12	1	0			50
SP70-1143 mt1	300972	118 (63)	5	10	97674-285433	45748-174192R	147
SP70-1143 mt2	144676	44 (27)	8	0			48
Chloroplasts							
SP80-3280 (Genomic)	141181						
SP80-3280 cp (gb)	141182	8	0	1			
SP80-3280 transcriptomic	141181	45 (0)	0	0			
IJ76-514	141176	26	2	5			
NCo310	141182	5	0	0			
RB72454	141181	7	0	0			
Q155	141181	0	0	0			
Q165	114181	2	0	0			
RB867515	141181	0	0	0			
SP70-1143	141181	2	0	0			
SP70-1143 (transcriptomic)	141181	43 (2)	0	0			
LCP85-384	141185	2	1	0			