Three founding ancestral genomes involved in the origin of sugarcane
Nicolas Pompidor, Carine Charron, Catherine Hervouet, Stéphanie Bocs, Gaëtan Droc, Ronan Rivallan, Aurore Manez, Therese Mitros, Kankshita Swaminathan, Jean-Christophe Glaszmann, et al.

To cite this version:

HAL Id: hal-03273957
https://hal.inrae.fr/hal-03273957
Submitted on 29 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Three founding ancestral genomes involved in the origin of sugarcane

Nicolas Pompidor1,2, Carine Charron1,3, Catherine Hervouet1,2, Stéphanie Bocs1,2, Gaëtan Droc1,2, Ronan Rivallan1,2, Aurore Manez1,2, Therese Mitros3, Kankshita Swaminathan4, Jean-Christophe Glaszmann1,2, Olivier Garsmeur1,2,§, and Angélique D’Hont1,2,§,∗

1CIRAD, UMR AGAP, F-34398 Montpellier, France, 2AGAP, Université de Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France, 3Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA and 4Department of Crop Sciences, University of Illinois, Urbana, IL 61801, USA

1These authors contributed equally to the work.
§These authors contributed equally to the work.
*For correspondence. E-mail dhont@cirad.fr

INTRODUCTION

Interspecific hybridization, sometimes accompanied by polyploidization, is an important evolutionary process in plants and is associated with the domestication and/or diversification of some major crops (e.g. banana (Simmonds 1962; Pernier et al., 2011), citrus (Wu et al., 2014), date palm (Flowers et al., 2019), rice (Santos et al., 2019) and wheat (McFadden and Sears, 1946).

Polyploids are generally divided into two categories: autopolyploids, which formed within a single species, and allopolyploids, which resulted from hybridization between two or more species. Autopolyploids are typically characterized by random association among homologous chromosomes during meiosis, leading to polysomic segregation, whereas allopolyploids have sets of homoeologous chromosomes that do not typically pair, leading to disomic segregation (Doyle and Egan, 2010). A continuum in the parental divergence of polyploids yields many intermediate situations (Stebbins, 1950; Barker et al., 2016).

Polyploidy and recurrent interspecific hybridizations complicate the reconstruction of phylogenetic relationships between genera and species, particularly in higher-order polyploids, which may have a complex history of multiple allo- and/or autopolyploidization events (Fortune et al., 2007; Tennessen et al., 2014; Triplet et al., 2014).

Sugarcane belongs to Saccharum sensu stricto, a genus composed exclusively of higher-order polyploids (≥4x) species. Despite its huge economic importance, the origin of sugarcane and the evolutionary history and taxonomy of the genus Saccharum (Poaceae; Andropogoneae) and its species are largely unresolved (Hodkinson et al., 2002; Welker et al., 2015). Several close genera (Erianthus section Rhipidium/Tripidiun, Miscanthus section Diandra, Narenga, Sclerochla) that can occasionally hybridize with Saccharum have been proposed by some authors to be involved in the origin of Saccharum and are referred to as the ‘Saccharum complex’ by breeders (reviewed

© The Author(s) 2021. Published by Oxford University Press on behalf of the Annals of Botany Company. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
by Daniels and Roach, 1987; Grivet et al., 2006). However, molecular data, although limited so far, do not support an important direct contribution of these genera to Saccharum but suggest a monophyletic origin of this genus (Grivet et al., 2004).

The subdivision of the genus Saccharum is a matter of debate (Irvine, 1999; Evans and Joshi, 2016), but a subdivision into six species is generally used by sugarcane technologists (Daniels and Roach, 1987; Grivet et al., 2006). Among them, two species are wild (S. robustum and S. spontaneum); they are well differentiated but for both species the taxonomic limit and evolutionary history have been a matter of controversy (reviewed by Daniels and Roach, 1987). Saccharum spontaneum (2n = 40–128) is a highly polymorphic species with an extensive distribution from Africa to Southeast Asia. It generally has thin stalks with no or very low sugar content. Saccharum robustum (2n = 60, 80 and up to 200) is most likely native to Southeast Asia, southeast to Sulawesi, and has long thick stalks with little or no sugar. Four ‘species’ exist only in cultivation (S. officinarum, S. barberi, S. sinense and S. edule) and are considered by some authors to be horticultural groups. The most popular scenario for sugarcane domestication, among sugarcane specialists, was first established by Brandes (1956). In this scenario, sugarcane originated in New Guinea from wild S. robustum by human selection possibly as much as 8000 years ago, and resulted in a series of clones accumulating sugar in the stalks identified by botanists as S. officinarum (2n = 80). These cultivars were transported by humans to the Asian continent, where they hybridized with local forms of the wild species S. spontaneum, giving rise to a new series of cultivars better adapted to subtropical environments and to the emergence of sugar manufacturing (Daniels and Daniels, 1976). They are called S. barberi for cultivars from India (2n = 81–124) and S. sinense for cultivars from China (2n = 116–120). The interspecific origin of these two groups of formerly cultivated sugarcane was demonstrated by molecular cytogenetics (D’Hont et al., 2002). Saccharum edule (2n = 60–122) is cultivated for its edible aborted inflorescence in subsistence gardens from New Guinea to Fiji and is believed to correspond to natural mutant clones from S. robustum (Grivet et al., 2006).

The origin of modern cultivars is well documented. They are all derived from a few interspecific hybridization events performed a century ago by breeders between the formerly cultivated groups S. officinarum and S. barberi and the wild S. spontaneum followed by backcrossing with S. officinarum (Arceneaux, 1968; Daniels and Roach, 1987). They are all high polyploids and aneuploids, with around 120 chromosomes, and molecular cytogenetics studies have highlighted that 75–85% of their chromosomes originated from S. officinarum and 15–25% from S. spontaneum, including some chromosomes derived from interspecific recombinations (D’Hont et al., 1996; Cuadrado et al., 2004; Piperidis et al., 2010; Huang et al., 2020; Piperidis and D’Hont, 2020). Recently, a reference sequence assembly of one mosaic basic genome of sugarcane has been produced (Garsmeur et al., 2018) as well as an assembly of a tetraploid S. spontaneum (Zhang et al., 2018). However, because of its extreme genome complexity, an assembly of the polyploid genome of a cultivar has not been obtained yet.

The objectives of the present study were to gain insight into the origin and architecture of the complex genome of modern sugarcane cultivars. For this, we analysed the sequences of all 12 hom(oe)ologous haplotypes (12 BAC clones) from two distinct genomic regions of a typical modern sugarcane cultivar. This allowed us to differentiate three groups of haplotypes. To investigate the origin of the haplotypes, we exploited sequence data from accessions representative of the diversity of Saccharum. We showed that two groups of haplotypes were contributed by S. officinarum and one group by S. spontaneum. These results suggested that three founding genomes were involved in the origin of the Saccharum genus and modern sugarcane cultivars.

MATERIALS AND METHODS

BAC sequencing and annotation

Ten BAC clones from the sugarcane cultivar R570 library developed by Tomkins et al. (1999) and identified by Jannoo et al. (2007) as corresponding to hom(oe)ologous chromosome segments bearing the Adhl gene were sequenced. Mate-pair libraries of ten BAC clones were produced and sequenced using the 454 method (FLX Titanium, Roche) and assembled with Newbler (Roche). Sequences were submitted to the EMBL database under the following accession numbers (BAC clone names in parentheses; Sh, Saccharum hybrid): HG531786 (Sh102M23), HG531788 (Sh111P05), HG531792 (Sh172H13), HG531793 (Sh182G15), HG531794 (Sh186P07), HG531797 (Sh192N12), HG531798 (Sh206M17), HG531799 (Sh209M19), HG531802 (Sh242M02) and HG531804 (Sh245F09). Two additional hom(oe)ologous BAC clones, Sh051L01 and Sh265O22 (accession numbers AM403006 and AM403007), were previously sequenced using the Sanger method (Jannoo et al., 2007).

Twelve BACs corresponding to hom(oe)ologous chromosome segments bearing the Rpal gene were identified and sequenced by de Setta et al. (2014). BAC sequences are available from GenBank under accession numbers KF184657 to KF184973.

For all BAC clone sequences, the structure (exon–intron) and putative function of genes were automatically predicted using the GNPAAnnot Community Annotation System (Guignon et al., 2012) available on the SouthGreen bioinformatics platform (https://www.southgreen.fr/). Gene predictions were manually curated using Artemis software as described in Garsmeur et al. (2011). Genes were numbered according to Jannoo et al. (2007) and de Setta et al. (2014) for the Adhl and Rpal regions, respectively. Large transposable elements (TEs) were annotated in the Adhl region as described in Garsmeur et al. (2011) and for the Rpal region the annotation from de Setta et al. (2014) was updated.

Identification of Miscanthus and Sorghum orthologous regions

Miscanthus sinensis and Sorghum bicolor orthologous regions were identified through BLASTN alignments of CDS sequences of all genes identified in the sugarcane BAC clones (16 and 6 genes for Adhl and Rpal regions, respectively) onto scaffolds of a preliminary genome assembly of Miscanthus sinensis and Sorghum bicolor genomes (assembly v3.0.1, available at https://phytozome-next.jgi.doe.gov/). BLASTN hits were filtered with
an e-value threshold of 1e−10, and for each region one *Sorghum* chromosome segment and two *Miscanthus* paralogous regions were identified. Genes were annotated as described above for sugarcane. The available *Sorghum* annotation (https://phytozome-next.jgi.doe.gov) was compared with the de novo annotations to help improvement of manual curation with Artemis software.

**Global sequence comparisons**

Sugarcane hom(oe)ologous BAC sequences and the *Miscanthus* and *Sorghum* orthologous regions were aligned using BLASTN. All alignments were inspected with Artemis Comparison Tools (Carver *et al.*, 2005).

**Phylogenetic analyses**

Hom(oe)ologous gene sequences, including exons and introns, were aligned with MAFFT (Katoh *et al.*, 2009). Maximum-likelihood phylogenetic trees for genes shared by most haplotypes were constructed using PhyML with the GTR evolution model and the SH-like aLRT branch support with 1000 bootstrap replicates (Guindon *et al.*, 2010).

Additional phylogenetic trees were constructed using the genomic segments shared between all hom(oe)ologous BAC sequences (from gene 6 to gene 7.5 for the *Adh1* region and from gene 1 to gene 5 for the *Rpa1* region) with an alignment-free method based on k-mer analysis using the AAF software (Fan *et al.*, 2015): hom(oe)ologous genomic segments were split into k-mers of 30 bp and homopolymeric k-mers were discarded. A pairwise distance matrix representing the number of k-mers that differed between hom(oe)ologous segments was used to reconstruct phylogenetic relationships.

Trees were visualized with Seaview (Gouy *et al.*, 2010) or FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

**Divergence times**

The number of substitutions per synonymous site (*Ks*) was calculated between all hom(oe)ologous gene pairs belonging to three groups of haplotypes (A, B and C). Protein sequences were aligned with Clustal W (Larkin *et al.*, 2007) and PAL2NAL (Suyama *et al.*, 2006) was used to reconstruct the multiple codon alignment based on the corresponding aligned protein sequences. The *Ks* values were calculated with the Nei–Gojobori method implemented in PAML (Yang, 2007). This process was performed using a script available at http://github.com/tanghaibao/bio-pipeline/tree/master/synonymous_calculation/. Divergence times were estimated using the formula 

\[ T = \text{average } K_s / (2 \times 6.5 \times 10^{-9}) \] 

(Cheng *et al.*, 1996).

**Sequence data from accessions representative of Saccharum species and relatives**

Two types of sequence data were used: whole-genome sequence (WGS) and targeted sequence capture data (Supplementary Data Table S1). Illumina paired-ends WGS data were available for two modern cultivars (including R570), 65 *Saccharum spontaneum* accessions and one *S. officinarum* accession (Garsmeur *et al.*, 2018; Zhang *et al.*, 2018). They represent on average a coverage of 7x of the total genome. In addition, pre-publication access to paired-end WGS data, with the Illumina NovaSeq S4 platform, for 16 *Saccharum* accessions (one modern cultivar, three *S. barberi*, two *S. officinarum*, one *S. robustum* and nine *S. spontaneum*) was provided by the Joint Genome Institute (JGI). They represent on average a coverage of 13x of the total genome.

Targeted sequence data were available for 304 *Saccharum* accessions and relatives (Yang *et al.*, 2019). This included sequences that mapped to 5914 sites of the *Rpa1* region but none that mapped to the *Rpa1* region. We produced a second set of targeted capture sequences for 36 accessions (6 modern cultivars, 10 *S. officinarum*, 15 *S. spontaneum*, 2 *S. barberi*, 1 *S. edule* and 2 *Miscanthus* accessions). Sequencing libraries were built with 1.5 μg of DNA by accessions using a protocol adapted from Kircher *et al.* (2012) and Meyer and Kircher (2010). DNAs were sheared to obtain an average of 300 bp on a Bioruptor® Standard (Catalogue No. UCD-200, Diagenode, Woburn, MA). Equal amounts of 16 genomic libraries were pooled to obtain at least 500 ng of DNA. Sequence capture by hybridization was performed on each library pool according to the manufacturer’s protocol for the myBaits® target capture kits (v3.02) with the custom oligonucleotide library designed by Arbor Biosciences. The regions targeted corresponded to 40 000 exons from sugarcane gene models annotated on the R570 reference sequence (Garsmeur *et al.*, 2018), including 17 476 sites from the two regions analysed in this study (*Adh1* and *Rpa1*).

**Single-nucleotide polymorphism identification**

Two subsets of read data were extracted from the WGS Illumina reads. The first one corresponded to reads that have a common k-mer (of size 20) with the BUSCO gene sets (Sepey *et al.*, 2019) present in the monoploid sugarcane reference sequence of Garsmeur *et al.* (2018). The second subset corresponded to reads that have a common k-mer (of size 20) with one of the gene exons of the *Adh1* or *Rpa1* region. *In silico* Illumina reads were generated from the R570 BAC sequences with the tool art_illumina (Huang *et al.*, 2012). For the targeted capture sequences, all the reads were used. All sequences were mapped on the monoploid sugarcane reference sequence of Garsmeur *et al.* (2018).

Single-nucleotide polymorphisms (SNPs) were identified as described by Garsmeur *et al.* (2018) (https://github.com/SouthGreenPlatform/VcfHunter/). This pipeline includes the mapping of sequence data onto the reference genome and identification and quality filtration of SNPs. For each accession and at each of the sites analysed, genotypes were determined if sequencing depth was at least 30. We coded genotypes as heterozygous if the variant occurred at least twice and at ≥4 % frequency. We coded genotypes as homozygous if no polymorphism was observed or if the variant occurred a single time and at a frequency of <1 % (which we considered to be a
potential sequencing error). Ambiguous cases that did not meet these criteria were coded as missing data. Note that a few SNPs detected in the R570 BACs were not detected in the R570 WGS data. They may represent sequencing errors, technical artefacts or small variations between the sequenced accessions that represent material that has been vegetatively propagated for many years.

**Multivariate analysis**

Factor analyses of distances table (AFTDs) were performed for the three distinct sets of data. For accessions for which WGS data were available, the SNPs detected in the BUSCO gene set were used. For the other accessions, SNPs detected in the targeted sequence data sets were used. AFTDs were performed with the DARwin program (Perrier and Jacquemoud-Collet, 2006), using dissimilarity matrices calculated on SNPs with an in-house program, vcf2dis.1.0.py, (https://github.com/SouthGreenPlatform/VcfHunter/blob/master/README.md).

**SNP origin analyses**

The origin, *S. officinarum*/*S. robustum* versus *S. spontaneum*, of the three groups of haplotypes/BACs (A, B and C) was determined using two complementary analyses. In the first analysis, SNPs specific to each of the three haplotype groups (A, B and C) were identified and their origin was inferred based on their presence/absence in the two germplasm pools (*S. officinarum*/*S. robustum* versus *S. spontaneum*). In the second analysis, SNPs specific to the two germplasm pools were identified and their distribution in the three haplotype groups (A, B and C) was analysed. These analyses were performed with two in-house python scripts (vcf2origin_AFB.py and vcf2origin_BFA.py respectively; https://github.com/SouthGreenPlatform/VcfHunter/blob/master/vcf2origin_AFB.py).

An SNP was considered specific to one group of haplotypes/ BACs (A, B or C) if (1) data were available for at least one haplotype for each group of haplotypes, and (2) the SNP was only found in one haplotype group. An SNP was used for origin analysis if (1) it was present in at least two accessions representative of one of the two germplasm pools and absent in all accessions of the other pool, and (2) if sequence data were available at its position for at least five *S. officinarum*/*S. robustum* accessions and five *S. spontaneum* accessions, among the accessions selected as representative of these germplasm pools. For the *Rpa1* region, since data were available for a lower number of accessions, only three *S. officinarum*/*S. robustum* were required. The same criteria were used when considering SNPs from the two germplasm pools (*S. officinarum*/*S. robustum* versus *S. spontaneum*) and examining their presence among the haplotypes/BACs. For the *Adh1* region, 45 *S. officinarum*, 8 *S. robustum* and 175 *S. spontaneum* accessions were used as representative of these species. For the *Rpa1* region, 13 *S. officinarum*, 1 *S. robustum* and 84 *S. spontaneum* were used as representative of these species (Supplementary Data Table S1). The position on the monoploid sugarcane reference sequence (Garsmeur et al., 2018) of a set of 31 832 SNPs identified as specific to *S. officinarum* versus *S. spontaneum* is shown in Supplementary Data Table S2.

**RESULTS**

**Comparison of two sets of 12 hom(oe)ologous haplotypes from modern cultivar R570**

Two hom(oe)ologous sets of haplotypes from cultivar R570 were analysed. The first set corresponded to a region bearing the *Adh1* gene located on sugarcane chromosome 1 (Garsmeur et al., 2018) and syntenic to *Sorghum* chromosome 1. This set contained 12 haplotypes, represented by 12 BAC clones. Sixteen genes with their corresponding allelic versions were annotated (Fig. 1, Supplementary Data Tables S3 and S4). The sequence that overlapped between pairs of hom(oe)ologous haplotypes varied from 35 to 113 kb, representing between 3 and 13 genes. One region of ~25 kb was shared by all hom(oe)ologous haplotypes and included three genes (genes 6–7.5).

The second set corresponded to a region bearing the *Rpa1* gene located on sugarcane chromosome 3 (Garsmeur et al., 2018) and syntenic to *Sorghum* chromosome 4. This set contained 12 haplotypes, represented by 12 BAC clones. Six genes with their corresponding allelic versions were annotated (Fig. 2, Supplementary Data Tables S3 and S4). The region shared by the 12 hom(oe)ologous haplotypes included five genes and represented around 35 kb (genes 1–5).

For both regions, the gene content and relative order were strictly conserved among all hom(oe)ologous sugarcane haplotypes. The percentage of nucleotide sequence identity was very high between all pairs of sugarcane hom(oe)ologous genes, with an average of 99.2 % for exons (ranging from 97.3 to 100 %) and 95.6 % for introns (ranging from 80.3 to 100 %).

The two regions were compared with their *Sorghum* and *Miscanthus* orthologous sequences. The gene content and order were also strictly conserved between sugarcane and the *Sorghum* orthologous sequences and the two ortho-paralogous *M. sinensis* sequences (Figs 1 and 2). Both sugarcane regions displayed a high level of nucleotide sequence conservation with *Miscanthus* and *Sorghum*, with an average of 96.5 % and 93.4 % for exons and 89.7 and 75.8 % for introns, respectively.

Large TEs were annotated in the *Adh1* and *Rpa1* regions, representing distinct classes of TEs [long terminal repeat (LTR) retrotransposons, non-LTR retrotransposons and DNA transposons] (Figs 1 and 2, Supplementary Data Table S4). Several TE insertion sites were conserved across hom(oe)ologous haplotypes in both regions.

**Phylogenetic relationships among hom(oe)ologous haplotypes distinguished three groups of haplotypes**

Phylogenetic relationships between hom(oe)ologous haplotypes were analysed based on (1) genomic regions shared by all haplotypes comprising genes 6–7.5 for the *Adh1* region and genes 1–5 for the *Rpa1* region (Figs 1 and 2) and (2) hom(oe)ologous copies of individual genes, including exons and introns (Figs 3 and 4).
For the Adh1 region, both analyses revealed three groups of haplotypes, with a major group (named A) that included nine haplotypes (Sh182G15, Sh209M19, Sh245F09, Sh206M17, Sh172H13, Sh242M02, Sh192N12, Sh111P05 and Sh102M23), a second group (named B) that included two haplotypes (Sh051L01 and Sh186P07) and a third group (named C) with a unique haplotype (Sh265O22).

For the Rpa1 region, both analyses also revealed three groups of haplotypes, in accordance with the result of de Setta et al. (2014), with a major group (A) that included ten haplotypes (Sh227O17, Sh232H22, Sh201D09, Sh239H20, Sh117K09, Sh142B14, Sh053L01, Sh101B12, Sh130J24 and Sh022O20) and two groups (B and C) each with a single haplotype (Sh051L01 and Sh186P07) and a third group (named C) with a unique haplotype (Sh265O22).

For the Adh1 region, both analyses revealed three groups of haplotypes, with a major group (named A) that included nine haplotypes (Sh182G15, Sh209M19, Sh245F09, Sh206M17, Sh172H13, Sh242M02, Sh192N12, Sh111P05 and Sh102M23), a second group (named B) that included two haplotypes (Sh051L01 and Sh186P07) and a third group (named C) with a unique haplotype (Sh265O22).

For the Rpa1 region, both analyses also revealed three groups of haplotypes, in accordance with the result of de Setta et al. (2014), with a major group (A) that included ten haplotypes (Sh227O17, Sh232H22, Sh201D09, Sh239H20, Sh117K09, Sh142B14, Sh053L01, Sh101B12, Sh130J24 and Sh022O20) and two groups (B and C) each with a single haplotype (Sh196O13 and Sh035B09, respectively). One exception was observed for gene 1 in haplotypes Sh227O17 and Sh232H22, which grouped separately from the other A haplotypes, suggesting that recombination may have occurred in these two haplotypes.

The two M. sinensis paralogues grouped generally together and always apart from the group of Saccharum hom(oe)ologous haplotypes.

TE insertion site conservation among hom(oe)ologous haplotypes reinforced the presence of three groups of haplotypes

In the Adh1 region, 15 TE insertion sites were conserved across two to nine hom(oe)ologous haplotypes (Fig. 1, Supplementary Data Table S4). Two TEs were shared only by the two haplotypes of group B (TE 8 and TE 9), nine TEs were conserved only across some or all available sequences for haplotypes of group A (TE 2, 4, 5, 7, 10, 11, 13, 14 and 15) and one TE (TE 3) was shared by haplotypes of groups A and B. The TE 1, TE 6 and TE 12 insertion sites were conserved across all haplotypes overlapping the corresponding regions. For TE 1, a 15-bp sequence corresponding to the short direct repeats of this TE was found in BAC Sh265O22. This suggested that TE 1 was present on the haplotype Sh265O22 but was removed by illegitimate recombination (Ma et al., 2004). The structure (complete versus fragment) was conserved among all other shared TEs, with two exceptions: TE 5 was found as a solo LTR in haplotypes Sh209M19 and Sh245F09, indicating that unequal homologous recombination between the two LTRs of
the complete TE 5 occurred; TE 6 was found fragmented on haplotypes Sh111P05 and Sh102M23.

In the Rpa1 region, in the interval that overlapped for the three groups of haplotypes, TE 4 was found on all haplotypes from group A, TE 6 on several haplotypes of group A and one TE (TE 5) was conserved on all haplotypes (Fig. 2).

Globally, the pattern of TE insertions among hom(oe)ologous haplotypes reinforced the distinction between the three groups of haplotypes, especially for the Adh1 region, and made it possible to position the TE insertion events on the phylogenetic tree (Figs 1 and 2).

Chronology of divergence between Saccharum homoeologous haplotype groups and with Miscanthus paralogous orthologues

Divergence times were estimated from synonymous substitution rates ($K_s$) for each gene of both the Adh1 and the Rpa1 R570 region. On average, divergence times within homologous haplotypes of group A and within haplotypes of group B was low (0.05–0.34 Mya). Divergence time between haplotypes from groups A and B, from groups B and C and from groups A and C were estimated to be 0.84, 1.23 and 1.29 Mya, respectively (Table 1).

Divergence time between Saccharum and Miscanthus lineages was estimated to be 5.7 Mya and the whole-genome duplication in Miscanthus to be 3.9 Mya.

Origin of the three groups of haplotypes coexisting in modern cultivars

To determine the origin of the three distinct groups of R570 haplotypes, we exploited sequence data from accessions representative of the Saccharum species. These data were aligned to the R570 sugarcane monoploid reference sequence and SNPs were identified.

Because hybridization can occur between the different Saccharum species and because mislabelling of accessions is frequent in collections, we performed multivariate analyses for each of the three sets of sequence data (the WGS data and the two sets of targeted sequence data) to analyse the structure
of the diversity within the *Saccharum* accessions and to select accessions representative of *S. spontaneum*, *S. officinarum* and *S. robustum*. For the data set from Yang *et al.* (2019), a preliminary multivariate analysis was performed to exclude accessions not belonging to the *Saccharum* genus. In this analysis, the first two axes clearly differentiated a large group...
of accessions labelled as *Saccharum* from scattered accessions that included accessions from closely related genera (*Miscanthus, Erianthus, Sorghum*) and several accessions probably mislabelled as *Saccharum* (Supplementary Data Fig. 1). These accessions were excluded and the remaining 272 *Saccharum* accessions were kept for further analysis.

In the multivariate analyses performed for each of the three sets of sequence data only with the *Saccharum* accessions, the first axis separated *S. officinarum* and *S. robustum* accessions from *S. spontaneum* accessions. Accessions belonging to *S. barberi,* *S. sinense* and modern cultivars were in an intermediate position in accordance with their interspecific origin (Supplementary Data Figs S2, S3 and S4). Multivariate analyses were then performed without these hybrid accessions. The first axis clearly separated *S. officinarum* and *S. robustum* accessions from *S. spontaneum* accessions (Fig. 5 and Supplementary Data Figs S3b and S4). *Saccharum spontaneum* accessions formed a very large group, with the second axis largely separating accessions from India from accessions from Indonesia in the first axis of a multivariate analysis performed with targeted capture sequences of 142 accessions (Fig. 5). A few accessions labelled as *S. robustum* and *S. spontaneum*...
had intermediate positions. They could represent hybrids between these groups and thus were excluded from the set of accessions chosen as representatives of the three species for the selection of specific SNPs. Due to their close proximity in the multivariate analysis, *S. robustum* and *S. officinarum* accessions were further considered as one germplasm pool while *S. spontaneum* accessions were considered as a second germplasm pool for the selection of SNPs specific to each of these two germplasm pools.

Two analyses were then performed with the selected accessions to study the origin of the three groups of R570 haplotypes/BACs (groups A, B and C). In the first analysis, we identified SNPs specific to each of the three groups of haplotypes/BACs in the *Adh1* and *Rpa1* regions (i.e. SNPs found only in all or some haplotypes from a single group). A total of 188 and 60 such SNPs were identified for the *Adh1* and *Rpa1* regions, respectively. Combining the two regions, 97 SNPs were found only in haplotype group A, 65 SNPs only in haplotype group B and 86 SNPs only in haplotype group C (Table 2 and Supplementary Data Tables S5 and S6). We then identified among the accessions selected as representative of *S. robustum/S. officinarum* and *S. spontaneum* which one had these SNPs. SNPs specific to a haplotype group in R570 but shared by both germplasm pools were interpreted as ancestral SNPs. All SNPs specific to haplotype groups A and B and present in only one germplasm pool were found only (with one exception) in the *S. officinarum/S. robustum* pool, suggesting that haplotypes A and B originated from this germplasm pool. This corresponded to 57 and 25 SNPs in the *Adh1* and *Rpa1* regions, respectively (Table 2). All SNPs specific to haplotype group C and found in only one germplasm pool were found only in the *S. spontaneum* pool, suggesting that haplotype C originated from this species. This corresponded to 28 and 16 SNPs in the *Adh1* and *Rpa1* regions, respectively (Table 2).

In the second analysis, from the sequences of the *Saccharum* representatives that mapped to the *Adh1* and *Rpa1* regions, we identified SNPs that were specific to each germplasm pool (i.e. SNPs found only in some or all *S. officinarum/S. robustum* accessions versus SNPs found only in some *S. spontaneum* accessions). A total of 96 and 44 such SNPs were identified for the *Adh1* and *Rpa1* regions, respectively (Table 3 and Supplementary Data Tables S7 and S8). We then identified

---

**Table 1.** Divergence time between the three groups of haplotypes (A, B and C) and between Saccharum and Miscanthus

<table>
<thead>
<tr>
<th></th>
<th>Ks Mya</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><em>Adh1</em></td>
<td>0.0043</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.007</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0112</td>
<td>0.0102</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0168</td>
<td>0.0166</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0173</td>
<td>0.0129</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0646</td>
<td>0.0874</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0480</td>
<td>0.0510</td>
</tr>
</tbody>
</table>

---

**Fig. 5.** First plane of a multivariate analysis separating accessions in two germplasm pools: one pool comprises *S. officinarum* (green) and *S. robustum* (blue) accessions, and one pool comprises *S. spontaneum* (red) accessions. Analysis is based on SNPs obtained from targeted capture sequences of 142 accessions (Yang et al., 2019). Accessions located between these two germplasm pools within the black rectangle may represent hybrids and were excluded for the selection of SNPs specific to these germplasm pools. Arrow points to accessions IN 84-088 and IN 84-089.
which of these SNPs were present in haplotype groups A, B and C. The SNPs found only in the *S. officinarum*/*S. robustum* accessions were exclusively (with one exception) found in R570 haplotype groups A and B while the SNPs found only in *S. spontaneum* accessions were exclusively found in R570 haplotype group C.

Both analyses clearly suggested that haplotype groups A and B originated from *S. officinarum* and *S. robustum*, while haplotype group C originated from *S. spontaneum*.

**Distribution of SNPs specific to the three groups of haplotypes in Saccharum**

The distribution of the 82 R570 SNPs specific to haplotype groups A and B that were found only in *S. officinarum*/*S. robustum* and of the 43 SNPs specific to group C found only in *S. spontaneum* was then analysed in the whole set of *Saccharum* accessions (Supplementary Data Table S1). We found SNPs specific to each of the three groups of haplotypes (A, B, C) in representatives of *S. barberi, S. sinense* and modern cultivars. These results are expected since they are hybrids between *S. officinarum* and *S. spontaneum* clones.

SNPs specific to each of the three groups of haplotypes (A, B, C) were also found in most of the accessions from *Yang et al.* (2019) that these authors re-classified as hybrids. Many of these accessions probably corresponded to mislabelled accessions, as already suggested by *Yang et al.* (2019).

**Table 2. Distribution of SNPs specific to each of the three groups of haplotypes in representatives of the Saccharum species in regions Adh1 and Rpa1**

<table>
<thead>
<tr>
<th></th>
<th><em>S. officinarum</em></th>
<th><em>S. robustum</em></th>
<th><em>S. spontaneum</em></th>
<th>All three species</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adh1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>29</td>
<td>0</td>
<td>44</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>28</td>
<td>1</td>
<td>19</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>27</td>
<td>40</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>57</td>
<td>28</td>
<td>103</td>
<td>188</td>
<td></td>
</tr>
<tr>
<td><strong>Rpa1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>15</td>
<td>0</td>
<td>9</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>0</td>
<td>7</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>16</td>
<td>3</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>16</td>
<td>19</td>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3. Distribution of SNPs specific to *S. officinarum*/*S. robustum* versus *S. spontaneum* in the three groups of haplotypes in regions Adh1 and Rpa1.**

<table>
<thead>
<tr>
<th></th>
<th>Group A</th>
<th>Group B</th>
<th>Group A and B</th>
<th>Group C</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adh1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>S. officinarum</em>/<em>S. robustum</em></td>
<td>29</td>
<td>28</td>
<td>4</td>
<td>0</td>
<td>61</td>
</tr>
<tr>
<td><em>S. spontaneum</em></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>29</td>
<td>4</td>
<td>27</td>
<td>89</td>
</tr>
<tr>
<td><strong>Rpa1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>S. officinarum</em>/<em>S. robustum</em></td>
<td>15</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td><em>S. spontaneum</em></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>10</td>
<td>2</td>
<td>16</td>
<td>43</td>
</tr>
</tbody>
</table>

The two *S. robustum* (IN 84-076 and IS 76-184) that we excluded as representative of *S. robustum* based on the multivariate analysis displayed one SNP specific to group C in addition to SNPs specific to groups A and B, which may indicate a hybrid status. The few *S. spontaneum* clones that we excluded as representative of *S. spontaneum* displayed SNPs specific to groups A and B in addition to SNPs specific to group C, suggesting a hybrid status. Two of them (IN 84-088 and IN 84-089 in Fig. 5) positioned very close in the multivariate analysis to the other *S. spontaneum* accessions from Southeast Asia (Indonesia). These two accessions carried only SNPs specific to groups B and C except for one SNP that was specific to group A (in region Adh1). However, data from one of the two BACs/haplotypes of group B was missing at this position; thus the SNP could also have been present on the missing BAC/haplotype, invalidating this position as specific for group A. In addition, this SNP was found in every *S. robustum*, *S. sinense* and *S. barberi* accession and modern cultivar and in 24 out of the 28 *S. officinarum* accessions. This could indicate that this SNP is an ancestral SNP common to haplotypes of group A and B but not C, rather than specific to group A. These results suggested that these two accessions (IN 84-089 and IN 84-088) could be hybrids between the B and C ancestral founding genomes.

**DISCUSSION**

We analysed 12 homoeologous haplotypes for two genomic regions in the genome of a typical modern sugarcane cultivar (R570). These regions belong to chromosomes 1 and 3, for which 12 chromosome copies were revealed with chromosome-specific oligo probes by FISH in cultivar R570 (*Piperidis and D’Hont*, 2020). This number is in the range of chromosome copy numbers expected for a modern cultivar (*Piperidis and D’Hont*, 2020). Our results revealed for both regions the existence of three groups of haplotypes, with a major group (A) being present in nine or ten copies and two minor groups (B, C) being present in one or two copies. The wild species are known in the *Saccharum* genus, *S. spontaneum* and *S. robustum*. The sweet domesticated canes, *S. officinarum*, are thought to have been domesticated from *S. robustum* (*Brandes, 1956*). SNPs specific to haplotypes A and B were found in representative accessions of *S. robustum* and *S. officinarum* but not in *S. spontaneum*. Conversely, SNPs specific for haplotype C were found in representative accessions of *S. spontaneum* but not in representatives from *S. robustum* and *S. officinarum*. The three groups of haplotypes were estimated to have diverged some
The detection of specific SNPs from two groups of haplotypes (A and B) in the genomes of both S. officinarum and S. robustum is consistent with the common view that S. officinarum has been domesticated from S. robustum (Brandes, 1956). Furthermore, our study clearly demonstrated that the foundation of S. robustum, and hence S. officinarum, is heterogeneous and that interspecific hybridization or allopolyploidization occurred in the evolutionary history of this taxon in addition to autopolyploidy. The much higher proportion of haplotype A compared with haplotype B observed in both studied regions of a modern cultivar suggests that the hybridization events involved transmission of unreduced gametes or hybridizations between autopolyploids with different ploidy levels and/or were followed by backcrosses with the A founder genome.

Recently, Zhang et al., 2019 contradicted the general assumption that S. officinarum was domesticated from S. robustum based on the divergence time of 0.38 Mya they estimated between assembled haplotypes from these two species. This divergence time is close to the one we obtained between haplotypes within group A (0.34 Mya). The presence of two founder genomes in the origin of these species, revealed by our study, may have complicated the interpretation of Zhang et al. (2019).

Specific SNPs from group C were found only in S. spontaneum. This species is highly polymorphic, with a large distribution range from Africa to Southeast Asia that overlaps with S. robustum, from Kalimantan Island to Papua New Guinea (Grivet et al., 2004). The high diversity of this species is illustrated by the results of the multivariate analysis, which showed a large group with two main subgroups, one from India and one from East Asia (Fig. 5). Specific SNPs from group C but not from groups A and B were found in clones from these two S. spontaneum subgroups. Several clones identified as S. spontaneum in collections were found in intermediate positions between these S. spontaneum groups and the S. robustum/S. officinarum pool. These clones may be natural hybrids between these species since they displayed SNPs specific to groups A, B and C. These clones were found in the multivariate analysis mainly between the S. officinarum/S. robustum pool and the S. spontaneum accessions from East Asia (Indonesia). This can be explained by the fact that their distribution overlaps in these regions (Grivet et al., 2006). Some of these S. spontaneum clones, based on pairwise genetic distance with modern cultivars, were suggested by Yang et al. (2019) to be the ones mainly involved in the origin of modern cultivars. Our analysis suggested instead that their genetic proximity with modern cultivars is linked to their hybrid status with S. officinarum or S. robustum as they bear A and B alleles specific to these species.

We found SNPs specific to each of the three groups of haplotypes (A, B, C) in representatives of S. barberi and S. sinense, in accordance with their proposed natural interspecific hybrid origin between S. officinarum and S. spontaneum (Brandes, 1956; Daniels and Roach, 1987; D’Hont et al., 2002). The three groups of SNPs were also found in all modern cultivars tested, which are all derived from a few interspecific hybridizations made by breeders a century ago between a few S. officinarum, S. spontaneum and S. barberi clones. These interspecific hybridizations were followed by backcrosses with S. officinarum to recover good agronomic performance. This process resulted in the reduction of the proportion of S. spontaneum chromosomes, which was estimated based on molecular cytotgenetics to represent between 15 and 25% of the chromosomes in modern cultivars (D’Hont et al., 1996; Cuadrado et al., 2004; Piperidis et al., 2010; Huang et al., 2020; Piperidis and D’Hont, 2020). More recently this proportion was shown to vary from one to four copies depending on the hom(oe)logy group in the few cultivars analysed (Piperidis and D’Hont, 2020).

The proportion of haplotypes A/B (11 haplotypes in the two regions) originating from S. officinarum versus haplotype C from S. spontaneum (one haplotype) observed in the two regions studied fits in this range.

This evolutionary model, implicating autopolyploid and allopolyploid/interspecific hybridization events, contradicts the assumption that S. officinarum has an autopolyploid origin (Garmsmeur et al., 2011; Vilela et al., 2017; Yang et al., 2017) and could explain the variable chromosome pairing affinity observed in Saccharum (Jannoo et al., 2007). A few genetic maps have been built and all of them are partial. However, they revealed some variation in paring affinity. Some preferential pairing has been observed in S. robustum and S. officinarum (Mudge et al., 1996; Edmé et al., 2006; Aitken et al., 2007) but not in S. spontaneum (Al-Janabi et al., 1993; da Silva et al., 1995; Alwala, 2008). This observation could suggest that S. spontaneum accessions are autopolyploids (from genome founder C) with polysomic pairing. This autopolyploidy was verified recently for one tetraploid clone, AP85-441 (haploid of SES 208), from which a genome sequence was assembled (Zhang et al., 2018). For S. robustum and S. officinarum, the observed pairing behaviour is compatible with a mix of allo- and autopolyploid origins with the coexistence of two groups of homologues (from genome founders A and B) resulting in preferential pairing (A versus B) but each pairing being in a polysomic manner. In a modern cultivar such as R570, based on our data and Piperidis and D’Hont (2020), it is probably not rare that haplotypes B and C are present in two copies. This could explain the strong preferential pairing that we have observed for some S. spontaneum and some S. officinarum chromosomes (Grivet et al., 1996; Hoarauf et al., 2001; Jannoo et al., 2004), while the other A haplotypes largely display polysomic inheritance. Occasional recombination between homoeologues may result in mosaic chromosome structures complicating the chromosome pairing picture observed (Jannoo et al., 2004).

Perfect collinearity and a very high level of gene structural conservation among hom(oe)ologous sugarcane chromosomes were observed, with an average divergence in coding sequence of <1% and all alleles being predicted to be functional (with one exception). The striking conservation of hom(oe)ologous genes that we observed confirmed and extended our previous results on two hom(oe)ologous BACs from the Adh1 region.
and seven homoeologous haplotypes from a region carrying the brown rust resistance (Brl1) gene (Jannoo et al., 2007; Garsmeur et al., 2011). One reason for this strikingly high level of gene conservation among homoeologous sugarcane haplotypes may be the relatively young age of the polyploidization event in Saccharum species, which may have given little time for the paralogous chromosomes to differentiate from each other. Another reason may be the high polyploidy and mixed allo- and autopolyploidy, with autopolyploidy restraining divergence through pairing and recombination between homologues and polysomic inheritance. The maintenance of a broad set of functional homoeologues could be involved in the remarkable productivity and phenotypic plasticity of sugarcane.

The comparison of the two sugarcane regions with Sorghum (2n = 2x = 20) orthologous regions and with the two ortho-paralogous Miscanthus (2n = 4x = 38) regions showed high gene synteny conservation. For both regions, gene phylogenetic analyses did not support the assumption that the allopolyploid event that arose around 3–4 Mya in the Miscanthus lineage after its divergence with Sorghum (2n = 2x) x (2n = 2x) productiv-ity and phenotypic plasticity of sugarcane. Functional homoeologues could be involved in the remarkable productivity and phenotypic plasticity of sugarcane.

ACKNOWLEDGEMENTS

We thank the South Green Bioinformatics platform (http://southgreen.cirad.fr) for providing a high-performance cluster, and the CIRAD Regional genotyping technology platform (GPTRG), Dr Dan S. Rokhsar, Dr Stephen Moose and Dr Anna Hale for sharing resources. We thank the JGI and collaborators for prepublication access to sugarcane diversity data for this study.

LITERATURE CITED


cultivars (Saccharum spp.) by molecular cytogenetics. Molecular & General Genetics 250: 405–413.


