Genomic insights into the recent chromosome reduction and polyploidization of complex autopolyploid sugarcane *S. spontaneum*

Jisen Zhang (zjisen@fafu.edu.cn)
Center for Genomics and Biotechnology, National Sugarcane Engineering Technology Research Center, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, College of Life Sciences https://orcid.org/0000-0003-1041-2757

Qing Zhang
Center for Genomics and Biotechnology of Haixia Institute of Science and Technology, Fujian Agriculture and Forestry University

Yiying Qi
Center for Genomics and Biotechnology, National Sugarcane Engineering Technology Research Center, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, College of Life Sciences

Haoran Pan
Fujian Provincial Laboratory of Haixia Applied Plant Systems Biology, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, 350002

Gang Wang
Fujian Agriculture and Forestry University

Xiuting Hua
FAFU and UIUC-SIB Joint Center for Genomics and Biotechnology, National Sugarcane Engineering Technology Research Center, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, F

Yongjun Wang
Fujian Provincial Laboratory of Haixia Applied Plant Systems Biology, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, 350002

Lianyu Lin
Fujian Provincial Laboratory of Haixia Applied Plant Systems Biology, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, 350002

Zhen Li
FAFU and UIUC-SIB Joint Center for Genomics and Biotechnology, National Sugarcane Engineering Technology Research Center, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, F

Yihan Li
Fujian Provincial Laboratory of Haixia Applied Plant Systems Biology, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, 350002

Panpan Ma
Fujian Provincial Laboratory of Haixia Applied Plant Systems Biology, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, 350002

Meijie Dou
Center for Genomics and Biotechnology, National Sugarcane Engineering Technology Research Center, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, College of Life Sciences

Yibin Wang
Fujian Agriculture and Forestry University

Hengbo Wang
Fujian Provincial Laboratory of Haixia Applied Plant Systems Biology, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, 350002

Xingtan Zhang
Fujian Provincial Laboratory of Haixia Applied Plant Systems Biology, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, 350002

Wei Yao
Guangxi key lab for sugarcane biology, Guangxi University, Nanning, Guangxi 530005, China

Yuntong Wang
Biomarker Technologies Corporation, Beijing, 101300, China

Xinlong Liu

Sugarcane Research Institute, Yunnan Academy of Agricultural Sciences, Yunnan Key Laboratory of Sugarcane Genetic Improvement
Maojun Wang
Huazhong Agricultural University https://orcid.org/0000-0002-4791-3742
Jianping Wang
University of Florida https://orcid.org/0000-0002-0259-1508
Zuhu Deng
National Sugarcane Engineering Technology Research Center, Fujian Agriculture and Forestry University, Fuzhou 350002, China
Qinghui Yang
Sugarcane Research Institute, Yunnan Agricultural University, Kunming, Yunnan Province, PR 650201, China
Baoshan Chen
Guangxi key lab for sugarcane biology, Guangxi University, Nanning, Guangxi 530005, China
Muqing Zhang
Guangxi Key Laboratory of Sugarcane Biology, Guangxi University
Haibao Tang
Fujian Agriculture and Forestry University https://orcid.org/0000-0002-3460-8570
Ray Ming
Department of Plant Biology, University of Illinois at Urbana-Champaign https://orcid.org/0000-0002-9417-5789

Article

Keywords: S. spontaneum, Saccharum, modern sugarcane cultivars

Posted Date: May 14th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-494691/v1

Version of Record: A version of this preprint was published at Nature Genetics on June 2nd, 2022. See the published version at https://doi.org/10.1038/s41588-022-01084-1.

Abstract

S. spontaneum is a founding *Saccharum* species that contributes stress resistance to the genetic background of modern sugarcane cultivars. Here, we have assembled the autopolyploid *S. spontaneum* Np-X genome with ancestral form into 40 pseudo-chromosomes in 10 homologous groups, revealing the recent chromosome reduction and polyploidization that occurred in *Saccharum*. The paleo-duplicated chromosomal pairs exhibit functional redundancy in *Saccharum* and underwent fission followed by fusion accompanied by centromeric spreading around 0.80 million years ago (Mya) before evolving into their current forms with basic chromosome numbers x = 9 and x = 8 in *S. spontaneum*, likely in a stepwise manner. WGDs occurred independently in *Saccharum* species around 1.5 Mya. Highly diverse chromatin structures exist among homologous chromosomes despite their high collinearity, and the re-structuring of NpChr5 and NpChr8 might have suppressed switching of chromatin structure from inactive to active. Resequencing of 116 sugarcane accessions elucidated that the *S. spontaneum* originated from North India and that the basic chromosome numbers x = 8, x = 9, and x = 10 originated independently, indicating that recent chromosome reduction rather than polyploidization has driven the adaptive evolution of *Saccharum*. Our study provides genomic resources and suggests new directions for accelerating sugarcane improvement and advances our knowledge of the evolution of auto-polyploids.

Introduction

Polyploidization is considered a major force in the evolution of plants, particularly in angiosperms, as fossil records indicated that up to 70% of the flowering plant species originated shortly after polyploidization[1]. Polyploids can generally be classified into two major categories: auto-polyploids and allo-polyploids[2-4]. Although auto-polyploids are considered to have arisen by whole genome duplication (WGD) of identical chromosome sets within a single species[5,6], the study of their genome evolution is hampered by the much lower availability of information than for allopolyploids due to the paucity of homologous chromosome-level genome assemblies in various auto-polyploid taxa.

Modern sugarcane (*Saccharum* spp., Poaceae) is a crucial crop with an economic value of 90 billion USD that provides 80% of the world's sugar and 40% of its ethanol yield, and its production by weight exceeds that of any other food crops such as rice, wheat, or maize (FAO, 2017). The sugar content of sugarcane juice varies from 1% to 24% Brix[7]. *S. spontaneum* is a founding *Saccharum* species that is widely distributed across a very large region from the Mediterranean to the Pacific. *S. spontaneum* is the genetic donor of stress tolerance to the genetic background of modern sugarcane hybrids, which has been a major breakthrough during sugarcane breeding. This species exhibits wide variation in chromosome numbers, which range from 2n = 40 to 2n = 128, and has ploidy levels ranging from 4x to 16x[7]. Moreover, *S. spontaneum* is particularly notable for the highest known degree of polyploidy within its genus as an hexadecaploid [7], and exhibits three basic chromosome numbers[8], x = 8, x = 9, or x = 10. The extensive variation in ploidy levels of *S. spontaneum* presents an extreme case for the study of the evolution of auto-polyploidy genomes in plants.

Recently, a mosaic monoploid genome for the modern sugarcane cultivar R570 (382 Mb)[9], a haploid genome assembly for *S. spontaneum* AP85-441 (3.13 Gb)[10], as well as a representative gene space assembly of the hybrid sugarcane variety SP80-3280 (4.26 Gb)[11] and an auto-octoploid *S. officinarum* assembly (Zhang *et al.*, under review) have been released. These genome assemblies have provided an opportunity to jointly characterize the evolutionary history of *Saccharum*. *S. spontaneum* Np-X (2n = 4x = 40), which grows along the Himalayas at over 1300 m above sea level, has the lowest total number of chromosomes among natural *Saccharum* accessions, and is the only accession with a basic chromosome number of x = 10 in *S. spontaneum* accession to our knowledge, so it likely represents the ancestral karyotype of *S. spontaneu*m and the intermediate karyotype for the evolution of *Saccharum*. Here, we have generated a high-quality auto-tetraploid genome assembly for *S. spontaneum* Np-X using circular consensus sequencing (CCS)[12] and performed resequencing of 102 *S. spontaneum* accessions with various ploidy levels and basic chromosome numbers. In addition to describing the very recent chromosome reduction and polyploidization events in *Saccharum*, our study presents a hereditary blueprint for the genomic basis of sugarcane biology and sheds new light on the evolution of auto-polyploids.

Results

An allele-defined genome for *S. spontaneum*

Circular consensus sequencing (CCS) can produce accurate reads from noisy individual subreads, representing a better tradeoff between read length and accuracy, thereby providing a better tool to tackle highly complex auto-polyploid genomes. We obtained a total of 52 Gb of PacBio CCS long reads from 909 Gb of raw sequence data and 417 Gb of Illumina short reads on the Sequel and HiSeq X platforms, respectively (Supplementary Table S1). The Canu (v1.9)[13] software package was used to perform initial *de novo* assembly of the *S. spontaneum* Np-X genome and yielded an initial contig-level assembly with an N50 of 405 Kb, a significant improvement relative to the previous *S. spontaneum* AP85-441 genome, which had an N50 of 45 Kb (Table 1 and Supplementary Note). The total size of this initial contig-level genome assembly was 2.76 Gb, which accounts for 97.53% of the Np-X genome size estimated by genome survey based on K-mers (2n = 4x, ~2.83Gb) (Table 1, Supplementary Fig. S1 and Supplementary Note). A total of 59.97 billion (98.73%) of 60.74 billion Illumina short reads could be aligned and covered 99.05% of the assembly (Supplementary Table S2), supporting that a high-quality contig-level genome had been obtained for further processing.

We then used ~105x of high-throughput chromosome conformation capture (Hi-C) data to scaffold the allele-aware chromosome-level auto-tetraploid *S. spontaneum* Np-X genome using ALLHiC[10,14] (Supplementary Fig. S2 and Table S3). A total of 2.74 Gb (98.96%) of contigs were anchored into 40 pseudo-chromosomes comprised of 10 homologous groups with four allelic chromosomes in each group (Fig. 1 and Supplementary Table S4). Comparison of homologous haplotypes A, B, C, and D revealed 8.20 million SNPs, 0.66 million insertion/deletion (InDels), and 4,033 structural variations (SVs) among 10.73 Mb of sequence with heterozygosity of 1.39% in the *S. spontaneum* Np-X genome (Supplementary Fig. S3 and Supplementary Table S5). A total of 241 (96.27%) complete gene models among the 248 ultra-conserved core eukaryotic genes (CEGs) in CEGMA and 1,389 (96.46%) among 1,440 conserved genes in BUSCO were recalled in our assembly (Supplementary Table S6, S7). The assembly allowed us to predict 30 potential centromeric regions with lengths ranging from 0.59 to 8.63 Mb and 43 telomere regions with monomer copy number from 104 to 1,218 along the 40 chromosomes (Supplementary Table S8, S9).

Allele-defined genome annotation

Allele-defined gene annotation is the key for characterizing autopolyploid genomes. Based on our homologous chromosome-level assembly, we annotated a total of 123,128 protein-coding gene models (Table 1 and Supplementary Table S4) and 93.5% (115,166) of these gene models were functionally annotated via searches of the NR, GO, KEGG, Swiss-Prot, and KOG databases (Supplementary Table S10 and Extended Data 1). Due to the polyploid nature of this genome, many annotated gene models are naturally allelic variations of the same gene. Analysis of allele definition showed that 35,830 gene models were well defined, including 8,645 (24.1%) genes with four alleles, 12,861 (35.9%) with three, 10,781 (30.1%) with two, and 3,543 (0.1%) with one, with an average of 3.4 alleles per gene (Table 1, Supplementary Table S11 and Extended Data 2). These results confirm that the chromosome-level *S. spontaneum* Np-X assembly and gene annotation are well-organized and allele-defined.

Among the gene models we defined, 99.38% (122,177) and 93.26% (114,656) of CEGs could be identified in the genomes of *S. spontaneum* AP85-441[10] (2n = 4x = 32) and sorghum (*Sorghum bicolor*[15])(2n = 2x = 20) (Extended Data 4), respectively, supporting that these two *S. spontaneum* accessions share more similar gene content, than do *S. spontaneum* and sorghum, as expected from their phylogenetic relationships. Comparison of the genome assemblies of *S. spontaneum* AP85-441[10], *Saccharum* hybrid R570[9], *Saccharum* hybrid SP3280[11], and sorghum[15] showed that, among 34,232 gene families, 1,934 (5%) were unique to *S. spontaneum* Np-X (Supplementary Fig. S5). The Np-X-specific genes were enriched in several Gene Ontology (GO) categories, including 'RNA-DNA hybrid ribonuclease activity', 'polysaccharide binding', 'glycolytic process', and 'defense response to fungus' (Supplementary Fig. S6). Our KEGG enrichment analysis indicated that some of these genes likely participate in glucose-6-phosphate isomerase pathway protein transport (Supplementary Fig. S7).

We identified a total of 1,590 Mb transposable elements (TEs) accounting for 57.52% of the assembled *S. spontaneum* Np-X genome (Supplementary Table S12), similar to their proportions in the genomes of AP85-441[10](56.02%), *S. officinarum* (57.79%), *Miscanthus*[16] (60.14%), and sorghum[15] (61.30%) (Supplementary Fig. S8a and Table S13). Long terminal repeat (LTR) retrotransposons account for 40.64% of the Np-X genome, including 11.47% Ty1-*copia* and 28.87% Ty3-*gypsy*, whereas LTRs account for about 39.64% of the AP85-441 genome, 41.59% of the *S. officinarum* genome, 43.55% of the *Miscanthus* genome, and 44.66% of the sorghum genome (Supplementary Fig. S8b and Supplementary Table S13).

The genome re-structuring in *S. spontaneum*

The genome of *S. spontaneum* Np-X shows strict synteny with that of sorghum except for a sorghum-specific inversion on SbChr04[10,15](Fig. 1c). The decrease in the basic chromosome number from 10 to 8 in *S. spontaneum* AP85-441 was caused by fission following by fusion of its ancestral sorghum chromosome homologs 5 and 8 as well as of rice chromosome homologs 11 and 12, which are paleo-duplicated chromosome pairs (PdCPs) that originated from the ρ event of the *Poaceae*[10,16,17]. However, the chromosome forms and numbers resulting from these events were not found in *S. spontaneum* Np-X (Fig. 1c and 1d). Relative to sorghum, two inversions occurred in *S. spontaneum* AP85-441 chromosomes APChr2AB and APChr7AB and three chromosomal fragments occurred in AP85-441 chromosome APChr6ABD[10], but these chromosomal variations were not detected in *S. spontaneum* Np-X or the related genus *Miscanthus* (Supplementary Fig. S4). These results indicated that chromosomal inversions occurred after the event resulting in the chromosome number decrease in *S. spontaneum* AP85-441, further confirming that *S. spontaneum* Np-X retained the chromosome forms of the last common ancestor (LCA) of *S. spontaneum*.

To analyze the scenario of the hypothesized rejoining of Chr5 and Chr8, we analyzed synteny between these *S. spontaneum* Np-X and *S. spontaneum* AP85-441 chromosomes (Fig. 2) and found that the recombination breakpoints in *S. spontaneum* Np-X were located on the centromeres of NpChr5 and NpChr8. We found that *S. spontaneum* AP85-441 APChr5 is comprised of the ancestral short arms of *S. spontaneum* Np-X NpChr5 and NpChr6, and APChr6 is comprised of the ancestral long arms of NpChr5 and NpChr7, and that centromere-specific sequences were retained and spread over longer distances in both APChr5 and APChr6 (Fig. 2b). However, although APChr2 is comprised of the ancestral long arms of NpChr8 and NpChr2, its centromere-specific sequences were retained from only ancestral NpChr2. In contrast, APChr7, which is composed of the ancestral short arms of NpChr8 and NpChr2, retained two centromere-specific sequences from the two Np-X chromosomes. Comparative analysis of the homologous genomic regions in Np-X and AP85-441 showed that a 6.2 Mb genomic region of NpChr5 containing 34 genes and a

6.1 Mb segment of NpChr8 containing 16 genes appears to have been lost in the centromeric regions of the AP85-441 genome (Fig. 2c, Supplementary Table S14 and S15).

To further investigate the concerted evolution of the PdCPs (NpChr5 and NpChr8) in *Saccharum*, close paralogous gene pairs were identified in Chr5 and Chr8 in Np-X as well as in the corresponding chromosomes of AP85-441 (Chr5A (57–89 Mb) and Chr6D (54–90 Mb) compared with Chr2 (98-125 Mb) and Chr7 (62–83 Mb)), *Miscanthus*[16] (Chr10 and Chr14), sorghum[15] (Sb05 and Sb08), and rice[18] (R11 and R12) (Fig. 2a). The numbers of pairs of gene conversions in each genome in descending order are: 848 pairs in rice, 213 in sorghum, 265 in Np-X, and 114 in AP85-441 (Supplementary Fig. S9, S10). Obviously, AP85-441 retained far fewer converted genes pairs than did Np-X. These phenomena could be the reason that re-structuring of the PdCPs results in the absence of homeologous genes during the concerted evolution of the chromosomal ancestors of NpChr5 and NpChr8.

Gene redundancy on PdCPs in *Saccharum*

The genes on NpChr05 and NpChr08 displayed lower expression levels than those located on the other chromosomes in the examined tissues, and a similar phenomenon was observed in the AP85-441 (Fig. 2e) and *S. officinarum* genomes (Supplementary Fig. S11). Moreover, the homologous regions between these two chromosomes displayed significantly higher average synonymous substitution rates (Ks, 0.036) within than did the homologous regions of other chromosomes (0.022) in Np-X (Fig. 2f), suggesting a more rapid evolution of the PdCPs.

Change in the chromatin status of PdCPs across the Poaceae

To investigate the evolution of the three-dimensional (3D) genome in the *S. spontaneum* Np-X auto-tetraploid, 1,937 million valid interaction read pairs were used to construct chromatin interaction maps for each set of homolog chromosomes at 500-Kb resolution. The frequencies of intra-chromosomal interactions displayed a rapid decrease with increasing linear distance (Supplementary Fig. S12). We found that 32.9–64.2% and 35.8–67.1% of regions on the homologous group chromosomes exist as compartment A (active regions) and compartment B (inactive regions), respectively (Supplementary Fig. S13 and S14), indicating that the distribution of chromatin compartment is disproportionate in the autopolyploid.

The A/B compartment changes were further investigated for the genes with a complete set of four alleles in the collinearity regions. We found that a total of 5,075 (59.9%) genes exhibited three conserved alleles and 2,240 (26.4%) genes exhibited two conserved alleles, while only 1,164 (13.7%) genes were conserved for all four alleles (Supplementary Fig. S15 and Supplementary Table S16). The GO term analysis of these three type genes indicated that the genes exhibited all four alleles conserved were mainly enriched in lignin catabolic process (GO:0046274; FDR < 6.14E-14), oxygen oxidoreductase activity (GO:0052716; FDR < 6.14E-14), multicellular organism development (GO:0007275; FDR < 3.60E-09) and ubiqultin-dependent protein catabolic process (GO:0006511; FDR < 3.60E-09), while the genes with two or three conserved were involved in transferase activity (GO:0016757; FDR < 4.15E-05), protein glycosylation (GO:0006486; FDR < 4.15E-05), superoxide metabolic process (GO:0006801; FDR < 6.71E-04) and peptidylprolyl cis-trans isomerase activity (GO:0003755; FDR < 9.05E-11) (Supplementary Fig. S16). At the chromosome levels, most of genes(68.8%, 59.8%, 81.6%, 56.5%, 84.6% 77.2% and 79.7%) in Chr1, 2, 4, 5, 6, 7 and 10 exhibited three conserved alleles($P < 1.0 \times 10^{-5}$, Fisher's exact test) and most genes (83.5%) in Chr9 exhibited two conserved alleles($P < 1.0 \times 10^{-5}$, Fisher's exact test), while most genes (55.1% and 42.8%) in only Chr3 and Chr8 exhibited all four conserved alleles($P < 1.0 \times 10^{-5}$, Fisher's exact test) (Supplementary Fig. S15 and Table S16). These results suggested that the A/B compartment diverged among sets of homologous chromosomes in this auto-polyploid species.

We further analyzed chromatin status in PdCPs of AP85-441[10], Np-X, sorghum[15], and rice[18]. In these four genomes, 44.8–47.5% and 52.5–55.2% of genomic regions exist in compartment A and in compartment B, respectively (Supplementary Table S17). We found that the genes of the homologous chromosomes of NpChr2/5/7/9 (74.0%–87.0%) reside mainly in the most conserved regions among the four genomes, while the genes (41.1–49.9%) that underwent switching from the B to A compartment reside mainly in the homologs of NpChr2/6 (Supplementary Fig. S17 and Fig. S18). It is noteworthy that NpChr5 and NpChr8 displayed the lowest degree of regions exhibiting B to A compartment switching among the homologous chromosome sets in Np-X compared to AP85-441 (Fig. 2d), and all these gene located in the regions of B to A compartment switching are higher expressed in AP85-441 than in Np-X (Supplementary Fig. S19), indicating that the re-structuring of NpChr5 and NpChr8 might have suppressed switching of chromatin status from inactive to active.

The recent polyploidization in *Saccharum*

Based on Ks estimations, *Saccharum* diverged from sorghum and *Miscanthus* ~6.4 Mya (Ks = 0.08) and ~4.0 Mya (Ks = 0.05), respectively. In *Saccharum*, *S. spontaneum* split from *S. officinarum* about 1.6 Mya (Ks=0.02), and the two *S. spontaneum* accessions, Np-X and AP85-441, separated at ~0.8 Mya (Ks=0.01) (Fig. 3a), demonstrating that chromosome reduction they underwent occurred very recently. Considering that x = 10 appears to be ancestral chromosome number in the Saccharinae-Sorghinae, the auto-octoploid of *Saccharum*, *S. spontaneum* SES208 (2n=8x=64) and *S. officinarum* (2n=8x=80) should have experienced two rounds of WGD, while the auto-tetraploid *S. spontaneum* Np-X experienced only one round of WGD after its divergence from sorghum (Fig. 3). Our Ks estimates reveal that the homologous chromosomes diverged at Ks = 0.01 in both Np-X and *S. officinarum*, and at Ks = 0.00 in AP85-441, indicating that the WGD of auto-octoploid *S. spontaneum* SES208 appeared to be more recent than those of both Np-X and *S. officinarum* (Fig. 3a).

TEs have long been considered as key genomic features that can trigger genome instability and are thus an important source of evidence for exploring the evolutionary history of genomes. In the four genomes (Np-X, AP85-441[10], *S. officinarum*, sorghum[15], and *Miscanthus*[16]) we examined, LTR retrotransposons appear to have undergone continuing and recent amplification bursts ranging from 0 to 2 Mya (Fig. 3b) The most recent LTR/*Gypsy* sequence was used as a reference to identify TE hits, and 10 distinct insertion peaks with identities ranging from 65% to 98% were detected in the four genomes (Fig. 3c and 3d). A Gaussian probability density function (GPDF) analysis estimated that the earliest TE insertion events (P1 and P3) occurred at ~2.0 Mya in *Saccharum* (Fig. 3d, 3e and Supplementary Fig. S20-22), which preceded the expected divergence time for *Saccharum*. The P4 insertion peak (71.0–72.2%) that occurred ~1.2 Mya could also be specifically identified in the *S. officinarum* genome. These results suggested that the divergence between *S. spontaneum* and *S. officinarum* occurred between 1.2 and 2.0 Mya, supporting the estimation of ~1.6 Mya based on the Ks values of orthologous gene pairs. The occurrence of the P6 peak (75.8–76.4%) ~0.89 Mya is found specifically in Np-X and AP85-441 and is consistent with the estimation of 0.8 Mya based on the analysis of orthologous gene pairs. Two TE insertion peaks with identities of 91.8–92.4% (P8, ~0.55 Mya) and 96.4–96.8 (P9, ~0.35 Mya) appeared specifically in Np-X genome, indicating that these TE insertion peaks occurred specifically in Np-X. Similarly, P2 (65.4–67.9%, ~1.9 Mya) and P5 (72.7–74.1%, ~1.0 Mya) TE insertion peaks were found in the sorghum genome, and a P7 (83.5–84.2%, ~0.6 Mya) TE insertion peak was found in *Miscanthus* (Fig. 3d, 3e and Supplementary Fig. S23, S24).

Genes related to the key characteristics of *Saccharum*

Given that Np-X exhibits the ancestral chromosome forms of *S. spontaneum*, comparative analysis of the genes in the three *Saccharum* and sorghum genomes may offer clues to the evolution of key agronomic characteristics of *Saccharum*. We thus analyzed the core gene families related to sugar accumulation, photosynthesis, and leaf width in *Saccharum*.

Genes encoding sugar metabolism enzymes and sugar transporters: We identified the key gene families related to sugar metabolism including sucrose phosphate synthase (SPS)[19,20], invertase (INV)[21], sucrose synthase (SUS)[22,23], and fructokinase (FRK)[24] from the three *Saccharum* genomes (Fig. 4a and Supplementary Table S18). Compared to sorghum, the number of INV genes had undergone expansion in *S. officinarum*, which exhibited high sugar content, but were relatively conserved in both of the *S. spontaneum* genomes, indicating that expansion of INV genes might be a prerequisite for evolution of sugarcane with high sugar content.

Sugar transporters are crucial for the allocation of sugars to sink and source tissues during the development of plants[25-27]. We identified a total of 119 genes that are likely members of sugar transporter superfamilies including PLT, STP/HXT, INT, VGT, pGlcT, SFP, TMT, and SWEET (Fig. 4a and Supplementary Table S18). The PLT, SUT, and SFP gene families expanded in the two *S. spontaneum* genomes, and the STP and VGT families showed more expansion in Np-X than in AP85-441, indicating that the expansion of the STP and VGT families in *S. spontaneum* Np-X could be a transitional intermediate state. In contrast, the SWEET gene family only expanded in *S. spontaneum* Np-X, suggesting that contraction of this gene family might have occurred in *S. spontaneum* AP85-441 after genome re-structuring. In addition, 19 STP and 23 PLT genes had expanded due to tandem duplications in AP85-441, while expansion of the PLT family was only observed in Np-X, indicating that the tandem duplications of STP genes in *Saccharum* occurred after speciation.

C4 photosynthesis pathway: The C4 photosynthesis pathway was discovered in sugarcane[28,29] and has been considered to have higher energy efficiency necessary for generating large biomass. We identified members of nine core gene families that participate in the photosynthesis pathway (CA, PEPC, PEPC-K, NAD-ME, PPDK, NADP-MDH, NADP-ME, PPEK-RP and PEPCK) in the three *Saccharum* genomes and that of sorghum (Fig. 4a and Supplementary Table S19). Compared to sorghum, gene expansions occurred in the NAD-ME gene family in both Np-X and AP85-441, but not in *S. officinarum* (Fig. 4a, Supplementary Fig. 25 and Table S19). *NpPEPC1*, *NpPEPC-k1*, *NpNADP-MDH2*, *NpNADP-ME2*, *NpPPDK1*, *NpPPDK-RP1*, *NpCA1*, and *NpCA2,* are preferentially expressed in leaves, suggesting that these eight genes might be related to C4 photosynthesis in Np-X, which is consistent with observations in AP85-441[10] and *S. officinarum* (Zhang *et al*., under review). However, with the exception of *NpPEPC1*, transcripts of other genes putatively involved in photosynthesis displayed much lower abundance (average TPM = 21.0) than in either AP85-441 (average TPM = 216.9) or *S. officinarum* (average TPM = 320.9) (Fig. 4b). Thus, it seems that the types of component genes involved in the C4 photosynthesis pathway in *Saccharum* might have converged as well as the regulation of their respective expression.

Narrow leaf genes: *S. officinarum* has much larger leaves than *S. spontaneum*, and among *S. spontaneum* varieties, Np-X has much smaller leave than AP85-411 (Fig. 1a). About six NARROW LEAF (NAL) genes controlling leaf width have been reported in rice[30-35], and among them, NAL1 was shown to have the strongest effect on leaf width. We identified 13 NAL genes in the three *Saccharum* genomes (Fig. 4a, 4c and Supplementary Fig. S26), and gene expression analysis showed NAL1 and NAL10 transcripts to be expressed at much lower abundance in Np-X than in either AP85-441 or LA-Purple, and both appear to be down-regulated compared to their expression in stem (Fig. 4c), suggesting these as candidate genes affecting leaf width.

The origination and independent polyploidization of *S. spontaneum*

We generated a total of 4,682 Gbp of resequencing reads for 102 *S. spontaneum* genomes and the genomes of 14 related species for population genetics analysis, using the reference *S. spontaneum* Np-X genome (Fig. 5a and Supplementary Table S20). A total of 3,345,380 high-confidence

variants including 3,140,400 SNPs and 204,980 InDels were identified in these data. Using the genomes of the 14 related species as outgroup (including Sorghum, Miscanthus, *S. officinarum* and *S. robustum*), principal component analysis (PCA) of SNPs showed substantial genetic diversity among *S. spontaneum* groups (Fig. 5b). PC1 (35.81%) and PC2 (15.64%) separated the *S. spontaneum* accessions from the outgroup accessions while PC1 *vs.* PC3 clearly divided the *S. spontaneum* accessions into four groups, which was consistent with our Maximum Likelihood (ML)-based phylogenetic analysis (Fig. 5a and 5c). The four groups displayed continuous geographic distribution from the Indian subcontinent to eastern and southern Asia. Group I accessions were primarily distributed in the Pakistan, India, and southwestern China (Himalayan region); Group II accessions were mainly distributed in southeastern China including Fujian and Taiwan provinces, and the Philippines; Group III accessions were mainly distributed in southern China including Yunnan, Guangdong, and Hainan provinces, and in Myanmar and Laos; and group IV accessions were primary distributed in Indonesia and India (Fig. 5b). It is noteworthy that the recorded progenitor *S. spontaneum* of modern sugarcane hybrids, Glagah, clustered into group IV.

Linkage Disequilibrium (LD) decays among the resequenced accessions were relatively low with an average rate of 0.086, supporting that these accessions had not experienced artificial selection (Supplementary Fig. S27). The π values of Groups were 0.24×10^{-3}, 0.31×10^{-3}, 0.29×10^{-3}, and 0.26×10^{-3}, respectively (Supplementary Fig. S28), which verifies the results of our LD analysis. Genetic differentiation values (F_{ST}) between any two of the four groups ranged from 0.047 to 0.084 (Supplementary Fig. S28), and each of the neighbor group pairs exhibited lower F_{ST} values in comparison to any two of the other four groups.

Further, we estimated admixture proportions and individual ancestry based on the SNP data set (Fig. 5d, Supplementary Fig. S29 and Fig. S30). In the admixture plot at $K = 5$ (Fig. 5d), each of the five groups exhibited distinct relative monophyletic ancestry matching our ML phylogenetic analysis, and further supports that a low level of genetic exchange between the four *S. spontaneum* groups due to their geographic isolation. Only weak gene flow was detected between Group I and Group II (P-value = 2.2e-308; F_statistic = 0.397) (Fig. 5e). Further, population structure analysis showed that 12 accessions in Group I had some Group II ancestry and that 11 accessions in Group II had some Group I ancestry (Fig. 5d), supporting the limited gene flow between the two groups. All of these findings indicate that the four *S. spontaneum* groups evolved relatively independently after they originated on the Indian subcontinent (Group I) and spread step by step to the regions where Group II, Group III, and Group IV are now distributed.

Consistent with the results of a previous study[10], our phylogeny showed that *S. spontaneum* accessions of diverse ploidy are clustered together within each group, confirming that the different ploidy levels of *S. spontaneum* were originated and diverged independently in each of these four groups.

The evolution of basic chromosome numbers in *S. spontaneum*

With the availability of the *S. spontaneum* Np-X genome (x = 10), basic chromosome numbers can be determined based on the coverage of mapping reads on the restructured chromosomes (NpChr5 and NpChr8), specifically by looking at read coverage across breakpoints for evidence of restructuring. Examples of x=9 typically has signs of restructuring around NpChr5, and not NpChr8 (Fig. 6a).

Totals of 4, 7, and 91 of the *S. spontaneum* accessions have basic chromosome numbers of x = 10, x = 9 and x = 8, respectively. It is noteworthy that the *S. spontaneum* accessions with x = 8 were distributed across all four groups, while the accessions with x = 10 and x = 9 formed a clade that was nested within Group I and were mainly located in Pakistan, northern India, and Tibet China (Fig. 5c). As accessions with basic chromosome numbers of x = 9 and x = 10 are only found in Group I, while those with basic chromosome numbers of x = 8 appear in all four groups of *S. spontaneum* accessions, we assumed that the fluid ploidy levels have evolved independently from ancestral progenitors with basic chromosome numbers of x = 8 in three groups: Group II, Group III, and Group IV. In addition, gene flow is undetected between any pairs of these three groups of accessions (Fig. 5e).

The population of accessions with a basic chromosome number of x = 8 shows much lower LD decays (an average of 0.043) than do those with basic chromosome numbers of x = 9 and x = 10 (an average of 0.127 and 0.256, respectively), supporting the notion that they had not undergone artificial selection. Tajima's D test suggested that the population of accessions with a basic chromosome number of x = 8 ($d = 1.225$) were mainly under balancing selection and a recent population contraction, while both those with basic chromosome numbers of x = 9 ($d = 0.191$) and x = 10 ($d = 0.17$) were under neutral selection (Supplementary Fig. S31). Moreover, accessions with a basic chromosome number of x = 8 exhibit a higher degree of nucleotide diversity ($\pi=3.80\times10^{-4}$) than the other two forms ($\pi = 3.26\times10^{-4}$ for accessions with basic chromosome number x = 9, and $\pi = 3.02\times10^{-4}$ for accessions with basic chromosome number x = 10) (Supplementary Fig. S32), indicating that balancing selection in the population of accessions with a basic chromosome number of x = 8 contributed to maintenance of high levels of polymorphism in those populations, perhaps also related to a large overall population size of x=8 species.

We explored the population demographic history of *S. spontaneum* using SNP data with a PMSC (Pairwise Sequentially Markovian Coalescence) model[31], which revealed that the *S. spontaneum* population experienced a prominent effective population (N_e) expansion ($N_e \sim 500{,}000$) around 12 to 14 Kya (thousand years ago) and a subsequent N_e contraction ($N_e \sim 10{,}000–60{,}000$) around 8 to 1.4 Kya (Fig. 6d). Interestingly, the time of the N_e expansion corresponded to that of the Marine Isotope Stage (MIS) 5e interglacial period, the warmest interval of a warming phase during which

the earth emerged from an extreme glacial phase[36-40], according to benthic oxygen isotope data. The time of the N_e contraction corresponded to the Younger Dryas cold event that occurred ~1.1–1.2 Kya during which the global climate changed dramatically[41-43]. The *S. spontaneum* population of accessions with a basic chromosome number of x = 10 diverged demographically from the *S. spontaneum* populations with basic chromosome numbers of x = 9 and x = 8 with a much smaller N_e in recent history.

The accessions of *S. spontaneum* with a basic chromosome number of x = 8 exhibit a high level of genetic diversity and are distributed over a broader geographical range than the other two forms of accessions with basic chromosome numbers of x = 9 and x = 10, indicating that the genome re-structuring likely contributed to the vigor and adaptation of *S. spontaneum* with a basic chromosome number of x = 8.

To investigate the genetic basis of likely fitness advantage of x = 8 in *S. spontaneum*, XP-CLR was used to detect natural selective sweeps. We identified selective sweep signatures for 25.7 Mb of the genome sequence with 338 candidate genes putative involved in the reduction of basic chromosome number from x = 10 to x = 8 (Supplementary Fig. S33 and Extended Data 5). The identified genes located within selective sweep regions are mainly enriched in the molecular function 'polysaccharide binding', suggesting divergence in the genetic control of related processes after the genome re-structuring in *S. spontaneum*. It is noteworthy that 13 of the genes found in the selective sweeps putatively encode glycosyltransferase, glycosyl hydrolases, STARCH SYNTHASE 1 (Npp.10B005850.1), sucrose-phosphate synthase (Npp.03C039020.1), cellulose synthase (Npp.01B008750.1), and SWEET (Npp.04A016530.1), indicating that divergence of genes involved in carbohydrate metabolism pathways likely occurred after the *S. spontaneum* genome re-structuring. Interestingly, four tandem FUTs (Xyloglucan fucosyltransferase) (Npp.04A015560.1, Npp.04A015590.1, Npp.04A015600.1, and Npp.04A015610.1) putatively involved in plant cell wall biosynthesis processes were also identified in the selective sweeps[44,45], which could provide a foundation for the morphological differentiation of the three populations. In addition, some gene functions related to stress responses were detected, e.g., LOT1(Npp.05D013900.1) which enhances ABA response and plant drought tolerance[46]; two genes (Npp.10B011680.1 and Npp.10B011690.1) putatively encoding PER3 and Npp.01B018240.1, which have been implicated in responses to low temperatures in plants[47,48]; two genes (Npp.06C016710.1 and Npp.06C016720.1) with functions related to cold tolerance and that are induced by drought and ABA; two genes (Npp.01B030420.1 and Npp.05D025320.1) encoding a putative ethylene response factor; and 16 genes, such as CYP81F (Npp.01B031390.1 and Npp.05D025440.1), which likely encode enzymes putatively involved in oxidation-reduction processes and might be related to tolerance of drought stress[49,50]. These results indicated that the natural genomic sweep regions were likely involved in polysaccharide metabolism and stress tolerance during the chromosome reduction process.

Discussion

The allele-defined genomes in our study clearly suggested that the founding *Saccharum* species are auto-polyploid, as a low level of divergence was detected among alleles based on Ks estimates (Ks = 0.00 or 0.01). Very recently, a whole-genome analysis demonstrated that *Miscanthus* originated from an allo-tetraploid event, confirming the independent WGD before the divergence of the *Miscanthus* and *Saccharum* genera[51], which disagrees with a previous study that suggested *Saccharum* shared an allopolyploid event with *Miscanthus*[16]. In *S. spontaneum*, the WGD of the x = 10 form (Ks = 0.01) likely occurred prior to that of the x = 8 form (Ks = 0.00) according to estimates of allelic divergence. Considering the x = 10 form to be the common ancestor, the WGD for x = 8 form must have been followed by reduction of the basic chromosome number from x = 10 to x = 8. Thus, we conclude that the different basic chromosome numbers of *S. spontaneum* originated from independent WGDs, and that the WGD in Np-X was close in time to the split of the x = 10 and x = 8 forms, as alleles within Np-X and Np-X/85-411 show similar levels of synonymous nucleotide divergence (Ks = 0.01). Likewise, the polyploidization of two founding *Saccharum* species, *S. spontaneum* and *S. officinarum,* likely occurred independently. *S. spontaneum* and *S. officinarum* are very recent auto-polyploids, and *S. spontaneum* became clearly separated from the rest of the *Saccharum* species (Fig. 3f). *S. spontaneum* and *S. officinarum* likely originated from a common ancestor with a basic chromosome number of x = 10 at ~1.6 Mya. Further, *S. robustum* distributes phylogenetically together with *S. officinarum* (Fig. 5c) and exhibits fewer interchromosomal rearrangements than does *S. officinarum*[52]. Thus, we assumed these three *Saccharum* species originated from the auto-polyploidization event.

Studies of recent genomic polyploidization are rare because genomic analysis of auto-polyploids is notoriously challenging. The auto-polyploid *Saccharum* genomes can bridge ancient and recent polyploidization events as the *Saccharum* genomes have experienced a recent WGD (less 0.80 Mya) and retained a set of PdCPs that originated from a much older WGD that occurred ~80 Mya affecting all cereal species. Ks estimates for the PdCPs NpChr5 and NpChr8 in *S. spontaneum* Np-X indicate relatively higher levels of nucleotide diversity between alleles (Fig. 2f). Interestingly, the NpChr5 centromere-specific sequences were found to have been retained and spread on both APChr5 and APChr6. Such functional redundancy likely resulted in genes on one PdCP, NpChr5, displaying the lowest transcript abundance among genes on the 10 *S. spontaneum* Np-X chromosome groups (Fig 2d). Ancestral NpChr5 split into two major segments and experienced translocations in the ancestors of NpChr06 and NpChr07 in all of the examined accessions with a basic chromosome number of x = 9 (Fig. 1d and Fig. 2ab). Therefore, we hypothesize that the ancestral NpChr5 split was the first step in the chromosome number reduction followed by rearrangement of ancestral NpChr8 to evolve into the x = 8 form. However, we cannot exclude the possibility that the ancestral NpChr8 split was the initial step for the chromosome reduction as we might not have identified all of the accessions with a basic chromosome number of x = 9. We might answer the questions raised by Kim *et al*[16] and consider the split of NpChr5 in x = 9 form as a stepping stone in the process of chromosome number reduction rather than assuming that the x = 9

form originated from crosses between the x = 8 and x = 10 forms. However, we can still hypothesize that *S. spontaneum* evolved after recent polyploidization followed by chromosome number reduction in diploids (Fig. 6e).

Although there is no significant global homologous genome dominance in *S. spontaneum*, over 60% of its genes display differences in allelic expression[10]. The auto-tetraploid *S. spontaneum* Np-X genome harbors four alleles for each gene, which results in broad redundancy of gene function in this species compared with diploid species. Our analysis of A/B compartment organization showed that only 1,164 (13.7%) of alleles sets were located in the regions in which all four alleles are conserved among the four homogenous groups, suggesting that auto-polyploidy might alleviate the functional redundancy caused by multiple alleles by dynamically regulating the spatial structure of chromatin in some regions. Similarly, in allo-tetraploid cotton, asymmetric A/B compartment switching was revealed between two subgenomes during polyploidization[53], which suggests that distinct spatial structures were retained in two sets of chromosomes even though they exhibit similar genome sequences. However, regions of NpChr5 and NpChr8 displayed the highest degrees of A to B compartment switching among the homologous chromosome sets in *S. spontaneum* Np-X compared to *S. spontaneum* AP85-441 (Fig. 2d). Thus, chromosome rearrangement might play as important roles in evolutionary dynamics as does polyploidization in auto-polyploids.

S. spontaneum Np-X plants have much narrower leaf widths and lower biomass compared to those of *S. spontaneum* AP85-441 and *S. officinarum.* Biomass production is mainly attributed to photosynthetic productivity in plants[54], and the recent polyploidization in *S. spontaneum* Np-X has likely affected only the regulation of its C4 photosynthesis pathway as the core gene family members in the photosynthesis pathway are conserved but their expression levels vary among the three species (Fig 4b). The candidate C4 genes display much lower transcript abundances in *S. spontaneum* Np-X than in either *S. spontaneum* AP85-441 or *S. officinarum*. Sugarcane performs an NADP-ME sub-type of photosynthesis, in which expression of the C4-type NpNADP-ME2 (TPM = 32.59), is much lower than that of APNADP-ME (TPM = 519.63) or SoNADP-ME (TPM = 277.49), reflecting the relatively lower efficiency of decarboxylation and lower activity of photosynthesis in Np-X. However, transcript abundances of the sugarcane homolog of the most important gene controlling leaf width identified in rice[55,56], NAL1, in descending order are SoNAL1, APNAL1, and NpNAL1 (Fig. 4c), which are correlated to the leaf width of the three species. Understandably, because of the recent WGD and chromosome reduction in *Saccharum*, the transcriptional regulation rather than gene gain/loss is likely a driving force in the strong environmental adaptability of *S. spontaneum,* and the divergence of transcriptional regulation of genes controlling the photosynthesis pathway and leaf width might form the genetic basis of the divergence in biomass productivity among *Saccharum* species.

The recent divergence of *Saccharum* less than 1.6 Mya has resulted in great variation in sugar content of its species. Among the gene families related to sugar metabolism, INV is the only gene family expanded in the high sugar-content species *S. officinarum* while the other genes examined were relatively conserved in the *Saccharum* genome. Whereas, in sugar transporter super families, with tandem duplication, the STP gene family had expanded in AP85-411, and the PLT family had expanded in both *S. spontaneum* varieties. The members of these gene families appear to have evolved independently in *Saccharum* species. Genes encoding proteins involved in sugar metabolism, such as SPS[57], FRK[24], SUSY[58], and sugar transporters such as SUT[59], SWEET[60] display differential expression between *S. spontaneum* and *S. officinarum*. Similar to our explorations of the genetic basis of biomass, the regulation of sugar accumulation *via* transcriptional regulation rather than the gain or loss of functional genes is the main avenue of exploration of the molecular mechanisms by which sugar accumulation diverged among *Saccharum* species.

S. spontaneum is widely distributed from Indonesia to New Guinea and Japan through the Indian subcontinent to the Mediterranean and Africa. A previous study suggested that *S. spontaneum* accessions could only be divided into two major groups[61] because a limited number of molecular markers in that study were not sufficient for exploring the genetic divergence of these large and highly polyploid genomes. Based on whole genome sequencing, our previous study showed that 64 *S. spontaneum* accessions could be divided into three distinct groups[10], and, the current study clustered the largest genetic collection thus far of 102 representative *S. spontaneum* accessions, into four groups. The phylogeny and geographical distribution indicated that the *S. spontaneum* was originated from the North of India, a biodiversity hotspots near the Himalayas with a climate that has been influenced by both east and south Asian monsoons[62], and then radiated along Middle East, East Asia and South-East Asia mainly with founder species of basic chromosome number of x=8. Obviously, *S. spontaneum* with x=8 is more adaptive to environment and facilitating diversification and radiations. However, the x = 10 form in *Saccharum* including *S. spontaneum*, *S. robustum*, and *S. officinarum* has only very limited geographical distributions. This phenomenon might be explained by the recent WGD only generating genetic redundancy rather than genetic variation, while both the recent genome re-structuring as well as genes involved in selective sweeps for functions related to polysaccharide metabolism and stress tolerance, have larger contributions to the adaptive evolution in *S. spontaneum*.

A *S. spontaneum* population with a basic chromosome number of x=10 and a small N_e diverged during a population contraction 8 Kya and might have had low adaptability to its environment. Around the same time, other *S. spontaneum* populations with basic chromosome numbers of x = 8 and x = 9 also underwent population contractions. Thus, the forms of *S. spontaneum* with basic chromosome numbers of x = 8 and x = 9 likely evolved from the form with a basic chromosome number of x = 10. The genome re-structuring that occurred within *S. spontaneum* has shown interesting functional and adaptive implications on the population scale.

Methods

Sampling and sequencing. Fresh young leaves were collected from individual *S. spontaneum* Np-X plants cultivated in a greenhouse kept at 25–30°C with 16 h light per day. Genomic DNA were isolated from these young leaves using a CTAB method[63]. The extracted DNA was then packaged with dry ice and sent to BioMarker company (Beijing, China) for construction of circular consensus sequencing (CCS) libraries and Illumina short read libraries that were subsequently sequenced on the PacBio Sequel and Illumina Hiseq platforms, respectively. Totals of ~52 Gb of PacBio HiFi reads and ~417 Gb of Illumina reads were generated for de novo assembly of the *S. spontaneum* Np-X genome.

Total RNA was extracted from mature stems and leaves of *S. spontaneum* Np-X using TRIzol (Invitrogen, Carlsbad, California). These two RNA samples were sent to Novogene (Beijing, China) for construction of transcriptome libraries and were sequenced on an Illumina platform.

Hi-C library construction and sequencing. Fresh tender leaves were collected from *S. spontaneum* Np-X plants and used to prepare chromatin cross-linked to DNA and fixed with formaldehyde as described previously[10]. The fixed samples were sent to BioMarker (Beijing, China) for Hi-C library construction and sequencing. Two Hi-C libraries were constructed after digestion with *Hind*III restriction endonuclease and sequenced on the Illumina HiSeq platform. Finally, a total of ~290 Gb (~105x) Hi-C reads were obtained for genome scaffolding (Supplementary Table S3).

Genome survey for genome size estimation. About 417 Gb of Illumina short reads from *S. spontaneum* Np-X was filtered using Trimmomatic[64] (v.0.36) software with default parameters. The clean reads were used to create Kmer set using Jellyfish[65] (v. 2.2.6). The genome size was estimated using K-mer (K=21) frequency-based methods with K-mer number/K-mer Depth. The genome size of *S. spontaneum* Np-X was estimated as 2,832.7 Mb with a heterozygosity rate of 1.14% based on a 21-Kmer distribution (Supplementary Fig. S1).

Genome assembly and scaffolding. About 52 Gb of CCS clean reads was used to assemble the *S. spontaneum* Np-X genome with Canu (v1.9)[13] using the optimized parameters “batOptions=-dg 3 -db 3 -dr 1 -ca 500 -cp 50”, Hicanu[66] with parameters batOptions="-eg 0.0 -sb 0.001 -dg 0 -db 3 -dr 0 -ca 2000 -cp 200", and Hifiasm[67] with parameter “-l0 -u”, respectively. Assemblies of 2.76 Gbp and N50 length of 405 Kbp by Canu (v1.9)[13], 2.97 Gb and an N50 length of 3,541 Kbp by HiCanu[66], and 2.89 Gb and an N50 length of 1,962 Kbp by HiFiasm[67] resulted. However, many chimeric contigs were identified in the latter two assemblies so that they could not be scaffolded with Hi-C using ALLHiC[14]. We attempted to solve the chimera problem using Hi-C interaction signals, but the scaffolds still aborted after chimera correction. After a comprehensive comparison of the results obtained by the above different assembly methods, the best-assembled Canu (v1.9)[13] version was finally chosen to serve as the reference genome for subsequent analysis.

A total of 290 Gb Hi-C reads were mapped on the contig-level assembly using BWA (v.0.7.15)[68] software with default parameters. The mapping results were pruned using an in-house script that generated a BAM file. All the contigs were reordered and scaffolded using the ALLHiC[10,14] pipeline. Finally, we generated a pseudo-chromosome assembled genome that included 40 chromosomes with a total length of 2,760 Mb and N50 of 67.73 Mb (Table 1). We mapped about ~60.7 Gb of Illumina short reads onto the assembled chromosome-level genome using Bowtie2[69] to estimate the quality of the assembly.

Gene annotation. Transcriptomic data were obtained from RNAs isolated from stem and leaf tissues of *S. spontaneum* Np-X. Trimmomatic[64](v. 0.36) software was used to further filter RNA-Seq data, and then HISAT2[70], which blocks duplicates, was used to align reads to the reference genome sequence. After reads were aligned, different coverage thresholds were set according to the sequencing depth of each aligned region to obtain reliable intron and optimal transcript information. Then we used TransDecoder (https://github.com/TransDecoder/TransDecoder/wiki) to predict the ORFs of optimal transcripts and define gene models. The optimal gene models were then screened and trained using AUGUSTUS[71] software. We chose protein sequences from closely related species of maize, sorghum, rice, sugarcane, and used them as input for Genewise (https://github.com/brewsci/homebrew-bio/blob/master/Formula/genewise.rb) software for gene prediction in the *S. spontaneum* Np-X genome. We next collected exon information for homologous proteins and transcripts, and collected intron information by comparing reads. AUGUSTUS[71] software was used to combine the above intron and exon information for gene prediction. The results of the above three methods were integrated, and then the PFAM database[72] was used for screening to obtain final gene prediction results. Finally, we used a Perl script to analyze the final assembled genome for eukaryotic genes and obtained a total of 123,128 high-confidence gene models.

Gene function annotation. Gene functions in the *S. spontaneum* Np-X genome were predicted using the best matches of the alignments as queries against the eggNOG[73], Nr[74] and Swiss-Prot[75] databases using BLASTP (E-value = 1e-5). The eggNOG, Nr, and Swiss-Prot databases were downloaded to our local server. Unigenes were then used to query the NCBI non-redundant nucleotide sequence (Nt) database using BLASTN with a cut-off E-value of 1e-5. We used a Perl script provided by EBI (https://www.ebi.ac.uk/) for InterPro annotation. The script sends the sequences to the official web server at InterProScan[76] for InterPro annotation and returns the results to our local server. In order to determine whether the proteins encoded by these genes might participate in any functional pathways, all gene models were aligned (E-value = 1e-5) to the KO database. The putative functions of the gene models were predicted and classified using the Clusters of Orthologous Group (COG) database and the Eukaryotic Orthologous Groups (KOG) database. The online KEGG Automatic Annotation Server was used to assign assembled sequences to Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways.

Genome collinearity and evaluation. Collinearity between the *S. spontaneum* Np-X, *S. spontaneum* AP85-441[10], sorghum[15], *Miscanthus*[51] and rice[77] genomes was analyzed using MCScanX[78] with default parameters, and all of the orthologous and paralogous gene pairs were identified based on synteny blocks. The paired synonymous substitution rates (Ks) were calculated using the Nei-Gojobori method (https://github.com/tanghaibao/bio-pipeline/blob/master/synonymous_calculation/synonymous_calc.py.)

Identification of A and B compartments. The principal component analysis method in the HiTC package (v. 1.28.0)[79] was applied to identify A and B compartments in *S. spontaneum* Np-X, *S. spontaneum* AP85-441, Sorghum and Rice. For this analysis, each chromosome was divided into consecutive 500-kb regions from the contact maps. Genomic regions belonging to the A compartments usually contain more genes than those corresponding to B compartments. To identify regions that had switched A/B compartment status during genome polyploidization, we considered only regions that showed changes of first principal component (PC1) values from positive to negative or vice versa in both biological replicates.

Repeat sequence annotation. Consensus transposable element (TE) sequences were generated using RepeatModeler with a combination of *de novo* and homology strategies including two *de novo* repeat-finding programs, RECON[80] and RepeatScout[81], which we imported into RepeatMasker (http://www.repeatmasker.org/) to identify and cluster repetitive elements. Tandem repeats were identified using the Tandem Repeat Finder (TRF) package[82], and unknown TEs were classified using TEclass[83]. Next, the outputs from the above processes were used to identify telomeres and centromeres. We also integrated results from LTR_FINDER[84] and LTRharvest[85] and removed false positives from the initial predictions using the LTR_retriever pipeline[86]. These LTRs were also classified as either intact or non-intact LTRs.

LTR burst time estimation. The most recent and longest LTR/*Gypsy* sequence was selected as the representative sequence for detecting additional TE hits in the genomes. Full-length and truncated LTRs with various lengths and identities were identified across genomes, and then each sequence (length = l) was divided into 30-bp units to determine the number of dots (TE hits) (n = l/30) with the same identity following a previously reported method[87]. All dots were used to generate a box plot according to their identities. Single peaks in the TEs identity distribution curves were separated for GPDF fitting and burst time calculation, and the average nucleotide substitution ratio (K) was defined as 2.58 standard deviations (σ). The formula t = K/r was used to calculate the TE burst time point for single peaks, where r refers to the nucleotide substitution rate for sugarcane species (6.5×10^{-9}).

Read mapping and variant calling. The raw paired-end reads were processed to remove adapter sequences using Trimmomatic (v. 0.36)[64]. The clean reads were then mapped onto the *S. spontaneum* Np-X genome using Bowtie2[69] with default parameters. The mapped reads were sorted using SAMtools (v. 1.3)[88] and duplicated reads were marked using Picard (https://github.com/broadinstitute/picard). We applied the Haplotypecaller function of GATK (v3.8;https://github.com/broadinstitute/gatk) to generate GVCF files for each accession followed by population variant calling using the GenotypeGVCFs tool in GATK. Then, more stringent filtering was applied to the raw variant set using VCFtools[89] with parameters '–max-missing 0.8 –maf 0.05 –max-alleles 2 –min-meanDP 4 –minQ 200'.

Population genomics analysis. The high-confidence set of variants was then used to perform population genetic analysis. Values of π and Tajima's D were calculated using VCFtools[89] in 500-Kb windows and 100-Kb steps based on the high-confidence filtered SNPs. PLINK[90] was used to perform principal component analysis and to transform the VCF file into a Plink binary file for input. The results of PCA were plotted using R[91]. ADMIXTURE[92] was then applied to infer population stratification among the 102 sugarcane accessions using the predefined number of genetic clusters K from 1 to 10. After the best value of K was calculated, the population structure of *S. spontaneum* was inferred using fastStructure[93] for K = 1 through K = 10. A maximum likelihood tree was constructed using RaxML[94] and the format conversion of the input file was performed using vcf2phylip (https://github.com/edgardomortiz/vcf2phylip). Finally, Figtree (https://github.com/rambaut/figtree/releases) was used to visualize the tree.

Analysis of population demographic history. We inferred a demographic history for *S. spontaneum* by applying the PSMC (Pairwise Sequentially Markovian Coalescence) model[35] to the complete diploid genome sequences. This method reconstructs the history of changes in population size over time using the distribution of the most recent common ancestor (tMRCA) between two alleles in an individual. Consensus sequences were obtained using SAMtools[88]. Bases with low sequencing depth (less than a third of the average depth) or high depth (twice the average depth) were masked. The analysis was performed using the following parameters: -N25 -t15 -r5 -p '4+25*2+4+6'.The mutation rate per generation per site was $6.5 \acute{} 10^{-9}$ and g = 2. PSMC modeling was performed using a bootstrapping approach, with sampling performed 100 times to estimate the variance of the simulated results.

Declarations

Data deposit

All raw sequencing data were deposited into the Sequence Read Archive (under Bioproject accession PRJNA721787). Genome assemblies and annotation files are stored in NCBI at the same accession and SGD (Sugarcane Genome Database) (http://sugarcane.zhangjisenlab.cn/sgd/html/index.html). Source data are provided with this paper.

Acknowledgements

This work was supported by the Science and Technology Planting Project of Guangdong Province (2019B020238001), the National Key Research and Development program (2018YFD1000104); the National High-tech R&D Program (2013AA100604); the National Natural Science Foundation of China (31201260, 31760413, and 31660420); the Science and Technology Major Project of Guangxi (AA17202025); the Fujian Provincial Department of Education (No. JA12082); the Natural Science Foundation of Fujian Province, China (Grant Number 2019J0102); The China Scholarship Council (201707877011); and the Scientific Research Foundation of the Graduate School of Fujian Agriculture and Forestry University Grant (324-1122yb050).

Author contributions

JZ conceived this genome project and coordinated research activities; JZ, QZ and RM designed the experiments; QZ, QY, WY, ZD, BC, MZ, JW and XL collected and generated sugarcane materials; QZ, GW, YQ, LL, XZ and HT assembled and annotated the genome; QZ, HP, YW, MW and HT analyzed the 3D genome; XH, YW, ZL, YL,PM and MD studied the genes relevant to the key characteristics in sugarcane; QZ, JZ and HT studied genome evolution; QZ, JZ, QY, HW contributed to the population genetic analysis; QZ and JZ wrote the manuscript.

References

1. Masterson, J. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* **264**, 421-4 (1994).
2. Kihara, H. & Ono, T. Chromosomenzahlen und systematische Gruppierung der Rumex-Arten. *Ztschrift Für Zellforschung Und Mikroskopische Anatomie* **4**, 475-481 (1926).
3. Clausen, J., Keck, D.D. & Hiesey, W.M. Plant evolution through amphiploidy and autoploidy, with examples from the Madiinae. (1945).
4. Ramsey, J. & Schemske, D.W. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Evol. Syst.* **29**, 467-501 (1998).
5. Jr, S.G. Types of polyploids; their classification and significance. *Adv. Genet.* **1**, 403-429 (1947).
6. Lewis & Walter, H. Polyploidy || Polyploidy in Species Populations. **10.1007/978-1-4613-3069-1**, 103-144 (1980).
7. Irvine, J.E. Saccharum species as horticultural classes. *Theor. Appl. Genet.* **98**, 186-194 (1999).
8. Piperidis, N. & D'Hont, A. Sugarcane genome architecture decrypted with chromosome-specific oligo probes. *Plant J.* **103**, 2039-2051 (2020).
9. Garsmeur, O. *et al.* A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nat. Commun.* **9**, 1-10 (2018).
10. Zhang, J. *et al.* Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. *Nat. Genet.* **50**, 1565-1573 (2018).
11. Souza, G.M. *et al.* Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. *GigaScience* **8**, giz129 (2019).
12. Wenger, A.M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155-1162 (2019).
13. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **10**, 1038/nbt.4277 (2018).
14. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833-845 (2019).
15. Paterson, A.H. *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-556 (2009).
16. Kim, C. *et al.* Comparative Analysis of Miscanthus and Saccharum Reveals a Shared Whole-Genome Duplication but Different Evolutionary Fates. *Plant Cell* **26**, 2420-2429 (2014).
17. Rice, C. & Sequencing, C. The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications. *BMC Biol.* **3**, 20 (2005).
18. Ouyang, S. *et al.* The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* **35**(Database issue), D883-7 (2007).
19. Galtier, N., Foyer, C.H., Huber, J., Voelker, T.A. & Huber, S.C. Effects of Elevated Sucrose-Phosphate Synthase Activity on Photosynthesis, Assimilate Partitioning, and Growth in Tomato (Lycopersicon esculentum var UC82B). *Plant Physiol.* **101**, 535-543 (1993).
20. Micallef, B.J. *et al.* Altered photosynthesis, flowering, and fruiting in transgenic tomato plants that have an increased capacity for sucrose synthesis. *Planta* **196**, 327-334 (1995).
21. Tang, G.Q. & Sturm, M.L. Antisense Repression of Vacuolar and Cell Wall Invertase in Transgenic Carrot Alters Early Plant Development and Sucrose Partitioning. *Plant Cell* **11**, 177-190 (1999).
22. Lingle, S.E. Sugar Metabolism during Growth and Development in Sugarcane Internodes. *Crop Sci.* **39**, 480-486 (1999).
23. Zhang, J., Arro, J., Chen, Y. & Ming, R. Haplotype analysis of sucrose synthase gene family in three Saccharumspecies. *BMC Genomics* **14**, 314 (2013).

24. Chen, Y. *et al.* Evolution and expression of the fructokinase gene family in Saccharum. *BMC Genomics* **18**, 1-15 (2017).
25. Sherson, S.M., Alford, H.L., Forbes, S.M., Wallace, G. & Smith, S.M. Roles of cell-wall invertases and monosaccharide transporters in the growth and development of Arabidopsis. *J. Exp. Bot.* **54**, 525-531 (2003).
26. Büttner, M. The monosaccharide transporter(-like) gene family in Arabidopsis. *FEBS Lett.* **581**, 2318-2324 (2007).
27. Kühn, C. & Grof, C.P. Sucrose transporters of higher plants. *Curr. Opin. Plant Biol.* **13**, 288-298 (2010).
28. Kortschak, H.P., Hartt, C.E. & Burr, G.O. Carbon Dioxide Fixation in Sugarcane Leaves. *Plant Physiol.* **40**, 209-213 (1965).
29. Hatch, M.D. & Slack, C.R. Photosynthesis by sugar-cane leaves. A new carboxylation reaction and the pathway of sugar formation. *Biochem. J.* **101**, 103-111 (1966).
30. Fujino, K. *et al.* NARROW LEAF 7 controls leaf shape mediated by auxin in rice. *Mol. Genet. Genomics* **279**, 499-507 (2008).
31. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496 (2011).
32. Fujita, D. *et al.* NAL1 allele from a rice landrace greatly increases yield in modern indica cultivars. *Proc. Natl. Acad. Sci. USA* **110**, 20431-20436 (2013).
33. Takai, T. *et al.* A natural variant of NAL1, selected in high-yield rice breeding programs, pleiotropically increases photosynthesis rate. *Sci. Rep.* **3**, 1-11 (2013).
34. Kubo, F.C., Yasui, Y., Kumamaru, T., Sato, Y. & Hirano, H.-Y. Genetic analysis of rice mutants responsible for narrow leaf phenotype and reduced vein number. *Genes & Genet. Syst.* **91**, 235-240 (2016).
35. Zhao, J., Luo, H., Jiang, Y., Yang, X. & Zha, R. Gene mapping for rice narrow leaf mutant Narrow leaf 11 (nal11). *Journal of Southern Agriculture* **48**, 1133-1138 (2017).
36. Imbrie, J. The orbital theory of Pleistocene climate: support from a revised chronology of the marine d18O record. *Milankovitch & Climate Part* **126** (1984).
37. Martinson, D.G. *et al.* Age dating and the orbital theory of the ice ages: Development of a high-resolution 0 to 300,000-year chronostratigraphy. *Quaternary Res.* **27**, 1-29 (1987).
38. Sarnthein, M. & Tiedemann, R. Younger Dryas-Style Cooling Events at Glacial Terminations I-VI at ODP Site 658: Associated benthic δ13C anomalies constrain meltwater hypothesis. *Paleoceanography* **5**, 1041-1055 (1990).
39. Szabo, Barney, J., Simmons & Kenneth, R. Thorium-230 ages of corals and duration of the last interglacial sea-level high stand on Oahu. *Science* **266**, 93-96. (1994).
40. Stirling, C.H., Esat, T.M., Mcculloch, M.T. & Lambeck, K. High-precision U-series dating of corals from Western Australia and implications for the timing and duration of the Last Interglacial. *Earth & Planetary Sci. Lett* **135**, 115-130 (1995).
41. Alley, R.B. *et al.* Holocene climatic instability: A prominent, widespread event 8200 yr ago. *Geology* **25**, 483 (1997).
42. Mayewski, P.A. *et al.* Major features and forcing of high-latitude northern hemisphere atmospheric circulation using a 110,000-year-long glaciochemical series. *J. Geophys. Res. Oceans* **102**, 26345-26366 (1997).
43. Severinghaus, J.P., Sowers, T., Brook, E.J., Alley, R.B. & Bender, M.L. Timing of abrupt climate change at the end of the Younger Dryas interval from thermally fractionated gases in polar ice. *Nature* **391**, 141-146 (1998).
44. Perrin, R.M. *et al.* Xyloglucan fucosyltransferase, an enzyme involved in plant cell wall biosynthesis. *Science* **284**, 1976-1979 (1999).
45. Vanzin, G.F. *et al.* The mur2 mutant of Arabidopsis thaliana lacks fucosylated xyloglucan because of a lesion in fucosyltransferase AtFUT1. *Proc. Natl. Acad. Sci. USA* **99**, 3340 (2002).
46. Qin, T., Tian, Q., Wang, G. & Xiong, L. LOWER TEMPERATURE 1 enhances ABA responses and plant drought tolerance by modulating the stability and localization of C2-domain ABA-related proteins in Arabidopsis. *Mol. Plant* **12**, 1243-1258 (2019).
47. Llorente, F., López-Cobollo, R.M., Catalá, R., Martínez-Zapater, J.M. & Salinas, J. A novel cold-inducible gene from Arabidopsis, RCI3, encodes a peroxidase that constitutes a component for stress tolerance. *Plant J.* **32**, 13-24 (2002).
48. Kim, M.J., Ciani, S. & Schachtman, D.P. A peroxidase contributes to ROS production during Arabidopsis root response to potassium deficiency. *Mol. Plant* **3**, 420-427 (2010).
49. Martínez-Ballesta, M. *et al.* The impact of the absence of aliphatic glucosinolates on water transport under salt stress in Arabidopsis thaliana. *Front. Plant Sci.* **6**, 524 (2015).
50. Essoh, A.P. *et al.* Exploring glucosinolates diversity in Brassicaceae: a genomic and chemical assessment for deciphering abiotic stress tolerance. *Plant Physiol. Biochem.* **150**, 151-161 (2020).
51. Mitros, T. *et al.* Genome biology of the paleotetraploid perennial biomass crop Miscanthus. *Nature Commun.* **11**, 1-11 (2020).
52. Zhang, J. *et al.* Recent polyploidization events in three Saccharum founding species. *Plant Biotechnol. J.* **17**, 264-274 (2019).
53. Wang, M. *et al.* Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. *Nat. Plants* **4**, 90 (2018).
54. Hofius, D. & Börnke, F.A.J. Chapter 13 - Photosynthesis, carbohydrate metabolism and source–sink relations. in *Potato Biology and Biotechnology* (eds. Vreugdenhil, D. *et al.*) 257-285 (Elsevier Science B.V., Amsterdam, 2007).

55. Qi, J. *et al.* Mutation of the Rice Narrow leaf1 Gene, Which Encodes a Novel Protein, Affects Vein Patterning and Polar Auxin Transport. *Plant Physiol.* **147**, 1947 (2008).
56. Takai, T. *et al.* A natural variant of NAL1, selected in high-yield rice breeding programs, pleiotropically increases photosynthesis rate. *Sci. Rep.* **3**, 2149 (2013).
57. Ma, P. *et al.* Comparative analysis of sucrose phosphate synthase (SPS) gene family between Saccharum officinarum and Saccharum spontaneum. *BMC Plant Biol.* **20**, 1-15 (2020).
58. Shi, Y. *et al.* Comparative analysis of SUS gene family between Saccharum officinarum and Saccharum spontaneum. *Trop. Plant Biol.* **12**, 174-185 (2019).
59. Zhang, Q. *et al.* Evolutionary expansion and functional divergence of sugar transporters in Saccharum (S. spontaneum and S. officinarum). *Plant J.* **105**, 884-906 (2020).
60. Hu, W. *et al.* New insights into the evolution and functional divergence of the SWEET family in Saccharum based on comparative genomics. *BMC Plant Biol.* **18**, 1-20 (2018).
61. Aitken, K. *et al.* Worldwide Genetic Diversity of the Wild Species Saccharum spontaneum and Level of Diversity Captured within Sugarcane Breeding Programs. *Crop Sci.* **58**, 218-229 (2018).
62. Jacques, F.M.B. *et al.* Quantitative reconstruction of the Late Miocene monsoon climates of southwest China: A case study of the Lincang flora from Yunnan Province. *Palaeogeogr., Palaeoclimatol., Palaeoecol.* **304**, 318-327 (2011).
63. Winnepenninckx, B., Backeljau, T. & De Wachter, R. Extraction of high molecular weight DNA from molluscs. *Trends Genet.* **9**, 407 (1993).
64. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
65. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770 (2011).
66. Nurk, S., Walenz, B.P., Rhie, A., Vollger, M.R. & Koren, S. HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291-1305 (2020).
67. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly with phased assembly graphs. *Nat. Biotechnol.* **18**, 170-175 (2020).
68. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).
69. Langdon, W.B. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min.* **8**, 1 (2015).
70. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357-360 (2015).
71. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309-W312 (2004).
72. Finn, R.D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222-30 (2014).
73. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309-D314 (2019).
74. Bleasby, A.J., Akrigg, D. & Attwood, T.K. OWL–a non-redundant composite protein sequence database. *Nucleic Acids Res.* **22**, 3574-3577 (1994).
75. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365-370 (2003).
76. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).
77. Ouyang, S. *et al.* The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* **35**, D883- D887 (2007).
78. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
79. Servant, N. *et al.* HiTC: exploration of high-throughput 'C' experiments. *Bioinformatics* **28**, 2843-2844 (2012).
80. Bao, Z. & Eddy, S.R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269-1276 (2002).
81. Price, A.L., Jones, N.C. & Pevzner, P.A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351-i358 (2005).
82. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573-580 (1999).
83. Abrusán, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329-1330 (2009).
84. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265-W268 (2007).
85. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
86. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of LTR retrotransposons. *BioRxiv*, 137141 (2017).

87. Huang, G. *et al.* Genome sequence of Gossypium herbaceum and genome updates of Gossypium arboreum and Gossypium hirsutum provide insights into cotton A-genome evolution. *Nat. Genet.* **52**, 516-524 (2020).
88. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
89. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-8 (2011).
90. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559-575 (2007).
91. Ginestet, C.J.J.o.t.R.S.S. ggplot2: Elegant Graphics for Data Analysis. **174**, 245-246 (2011).
92. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655-1664 (2009).
93. Raj, A., Stephens, M. & Pritchard, J.K. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics* **197**, 573 (2014).
94. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

Table

Table 1. Statistics of the *S. spontaneum* Np-X and *S. spontaneum* AP85-441 genome assemblies.

	S. spontaneum AP85-441 (2n = 4x = 32)	*S. spontaneum* Np-X (2n = 4x = 40)
Total assembly size of contigs (bp)	3,134,067,138	2,759,716,890
Number of contigs	91,867	14,872
Maximum length (bp)	400,129	3,900,625
Average length (bp)	34,095	185,564
N50 contig length (bp)	45,041	405,181
N90 contig length (bp)	17,099	83,183
Total assembly size of scaffolds (bp)	3,140,615,268	2,765,453,113
Number of scaffolds	15,768	716
Maximum length (bp)	126,636,275	104,759,444
Average length (bp)	199,176	3,862,364
N50 scaffold length (bp)	91,359,291	67,739,541
N90 scaffold length (bp)	61,676,436	54,872,887
Number of chromosomes	32	40
Number of gaps	76,099	14,162
Number of gene alleles	112,788	123,128
Number of genes	35,525	35,830
Number of genes with 4 alleles	4,289	8,645
Number of genes with 3 alleles	9,792	12,861
Number of genes with 2 alleles	14,797	10,781
Number of genes with 1 alleles	6,647	3,543
Average gene length (bp)	3,648	3,412
Average exon number	4.92	4.72

Figures

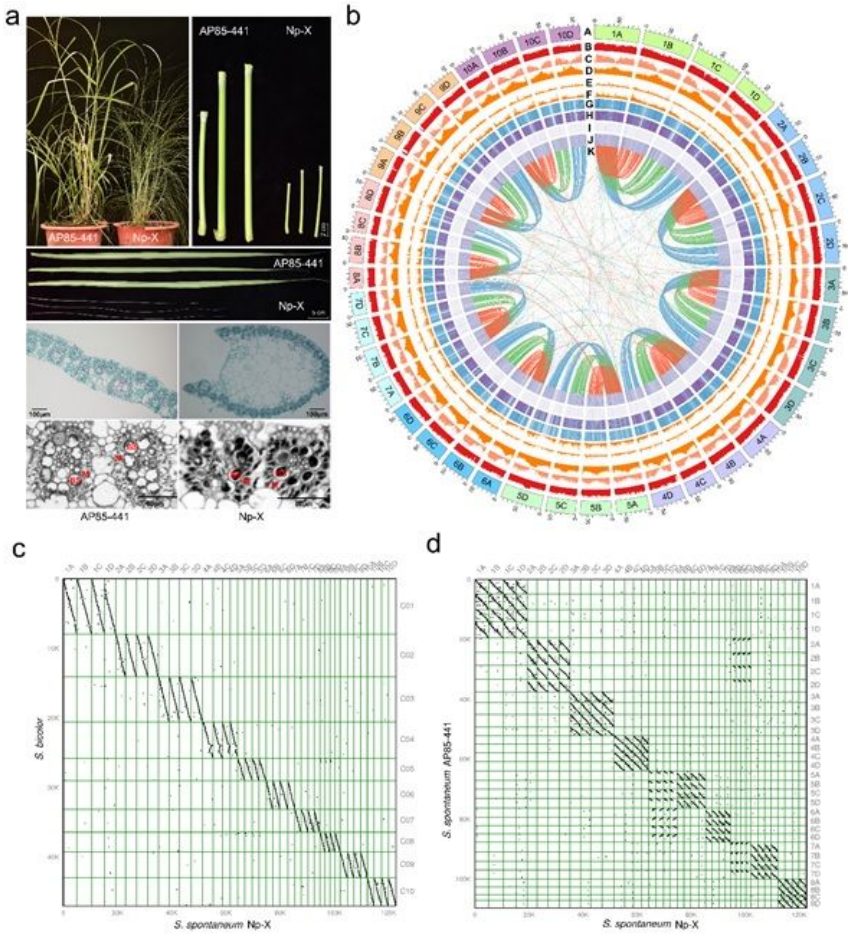

Figure 1

The morphological and genomic features of S. spontaneum Np-X. (a) Comparison of the morphological features of S. spontaneum Np-X and S. spontaneum AP85-441, from left to right and top to bottom indicating accession name, internode (internode 2, 4, 6 for Np-X, and 3, 6, 9 for AP85-441), leaf (leaf 1, 2, 3), cross section shape (long rectangular for S. spontaneum AP85-441 and ellipse for S. spontaneum Np-X) of leaves, and anatomical structure of leaves. BS, bundle sheath cell; M, mesophyll cell. (b) The genomic features of S. spontaneum Np-X. The tracks indicate (from outermost to innermost): (A) All 40 Pseudo-chromosomes in Mb; (B) GC content; (C) Gene density; (D) TE density; (E) SNP density; (F) InDel density; (G) distribution of Tajima's D values; (H) p values; (I) the transcriptomes of leaves and (J) stems; (K) and links in the inner circle where red ribbons indicate matching synteny blocks between ChrXA and ChrXB, green ribbons indicate synteny between ChrXA and ChrXC, blue ribbons indicate synteny between ChrXA and ChrXD, where X represents 1–10. (c) Alignment of S. spontaneum Np-X chromosomes with sorghum and S. spontaneum AP85-441 chromosomes (d). A set of four homologous S. spontaneum Np-X chromosomes aligned to a single sorghum chromosome (c) and to a set of four homologous S. spontaneum AP85-441 chromosomes (d), respectively.

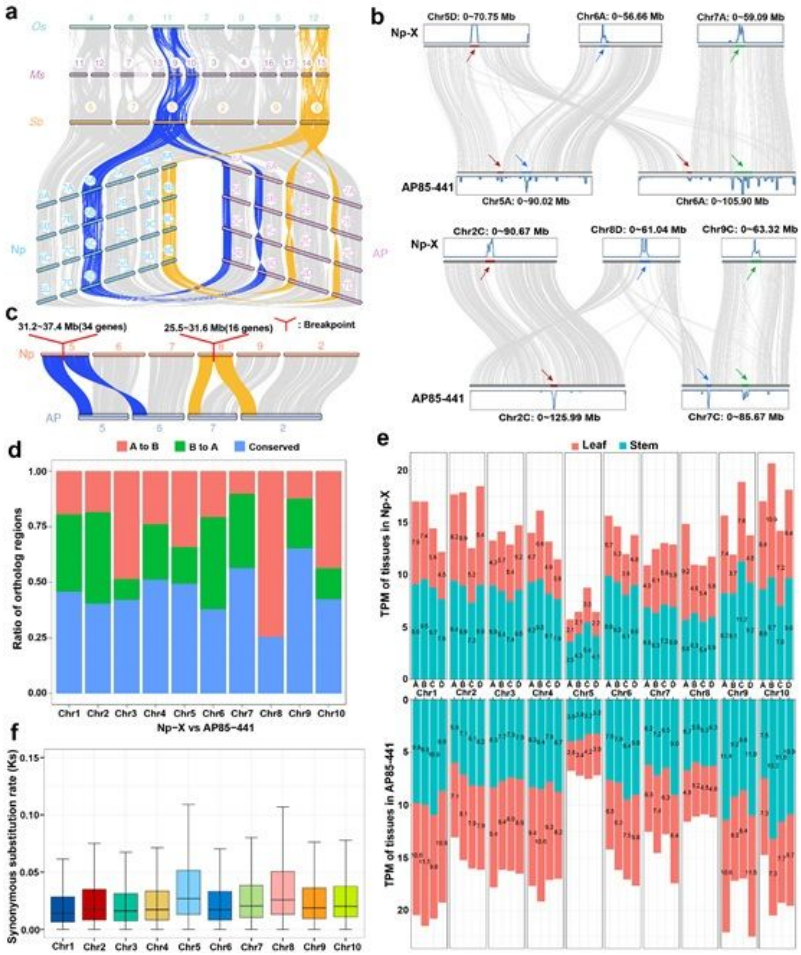

Figure 2

The evolution of paleo-duplicated chromosome pairs. (a) Characteristics of the evolution of ancestral Chr5 and Chr8 in the Poaceae. Genomic collinearity blocks are linked by gray lines. The synteny blocks in ancestral NpChr5 and NpChr8 between different species are highlighted as blue and yellow links, respectively. (b) The collinearity blocks between chromosomes are indicated with gray lines, and the blue line with peaks represents the density of centromere-specific sequences. Each color of arrows and line indicates the corresponding centromeric region. (c) The lost genes and segment sizes at the recombination breakpoints are shown as red lines. The synteny blocks of Chr5 and Chr8 in S. spontaneum Np-X compared with the corresponding regions in S. spontaneum AP85-441 are highlighted with blue and yellow lines, respectively. (d) Recombination of higher chromatin structure in the genomes of S. spontaneum Np-X and S. spontaneum AP85-441. The red, green, and blue bars show A-to-B compartments, B-to-A compartment switching, and conserved compartments during species evolution, respectively. (e) Total allelic expression level from the chromosomes of S. spontaneum Np-X and S. spontaneum AP85-441 in mature leaves and stems compared to S. spontaneum Np-X as reference. A–D represent four homologous chromosomes in each group. (f) Allelic synonymous base substitution rate (Ks). The distribution of allelic Ks in each homologous group is represented as a box plot.

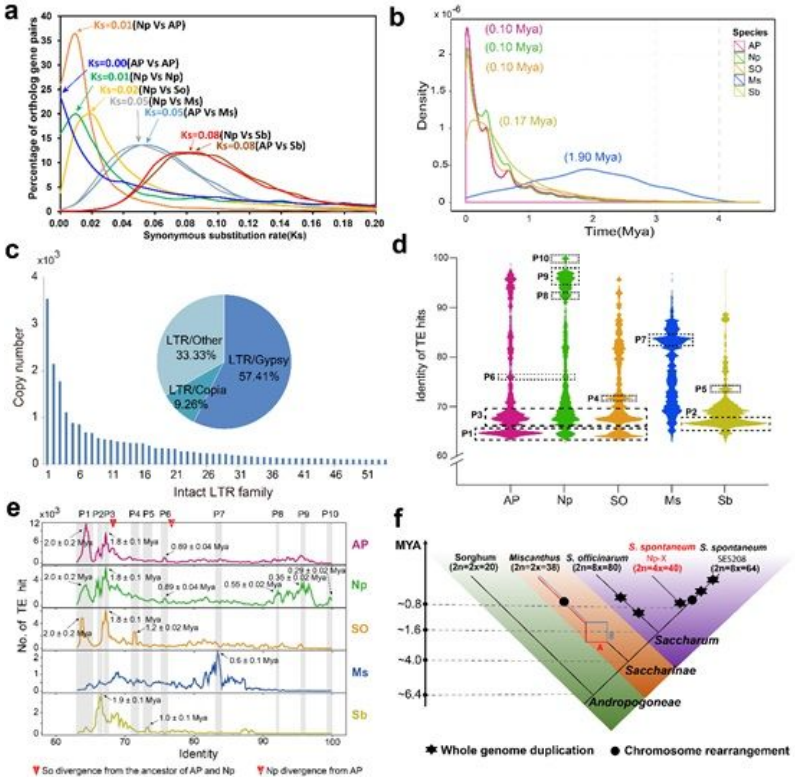

Figure 3

The evolution of Saccharum. (a) Distributions of synonymous substitution rate (Ks) between the representative species in Andropogoneae. Peak values for each comparison are displayed in corresponding colors. (b) Analysis of intact LTR insertion time in five species. The horizontal axis represents the insertion time of intact LTR-RTs, and the vertical axis represents the density of the number of intact LTR-RTs. (c) Classification of intact LTR-RTs in the S. spontaneum Np-X genome. LTR families with more than 100 copies are shown. (d) Sequence identity distribution of TE hits represented in a swarm plot. The most recent and longest LTR/Gypsy sequence among LTR families was chosen as the representative sequence for detecting additional TE hits in the genomes. A total of 85,022 dots (TE hits) in S. spontaneum AP85-441, 88,087 in S. spontaneum Np-X, 53,927 in S. officinarum, 28,135 in Miscanthus, and 22,849 in sorghum were identified. The areas in the box (P1–P10) represent 10 distinct burst events in different genomes. (e) Number of TE hits with the representative sequence as the query sequence and their associated identity values. The calculated burst time based on GPDF fitting of each peak is indicated at the arrow. The 10 peaks, P1–P10, defined in (d) are highlighted as shaded gray columns. The first red triangle represents the divergence between S. spontaneum and S. officinarum occurred at ~1.6 Mya and the second one represents the divergence between S. spontaneum Np-X and S. spontaneum AP85-441 at ~0.8 Mya. (f) A schematic map of the evolution of Saccharum. The stars and circles represent the WGD and chromosome rearrangement events, respectively.

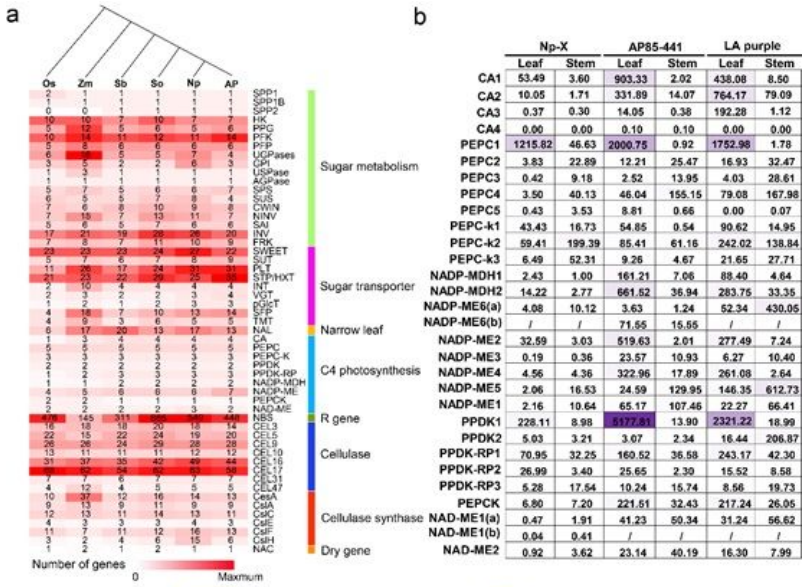

Figure 4

The gene families associated with crucial biological traits sugarcane. (a) Expanded gene families in sugarcane that could control crucial biological traits in S. spontaneum Np-X compared with those of rice, maize, sorghum, S. officinarum, and S. spontaneum AP85-441. The number of genes in each family are indicated with a heat map. (b) The expression profiles of gene families involved in C4 photosynthesis and (c) Narrow Leaf (NAL) genes in mature leaves and stems of S. spontaneum Np-X compared with those of S. spontaneum AP85-441 and S. officinarum. TPM from low to high is indicated by the shading of purple from light to dark. The green triangles from left to right represent leaf widths from the widest in S. officinarum to the narrowest in S. spontaneum Np-X.

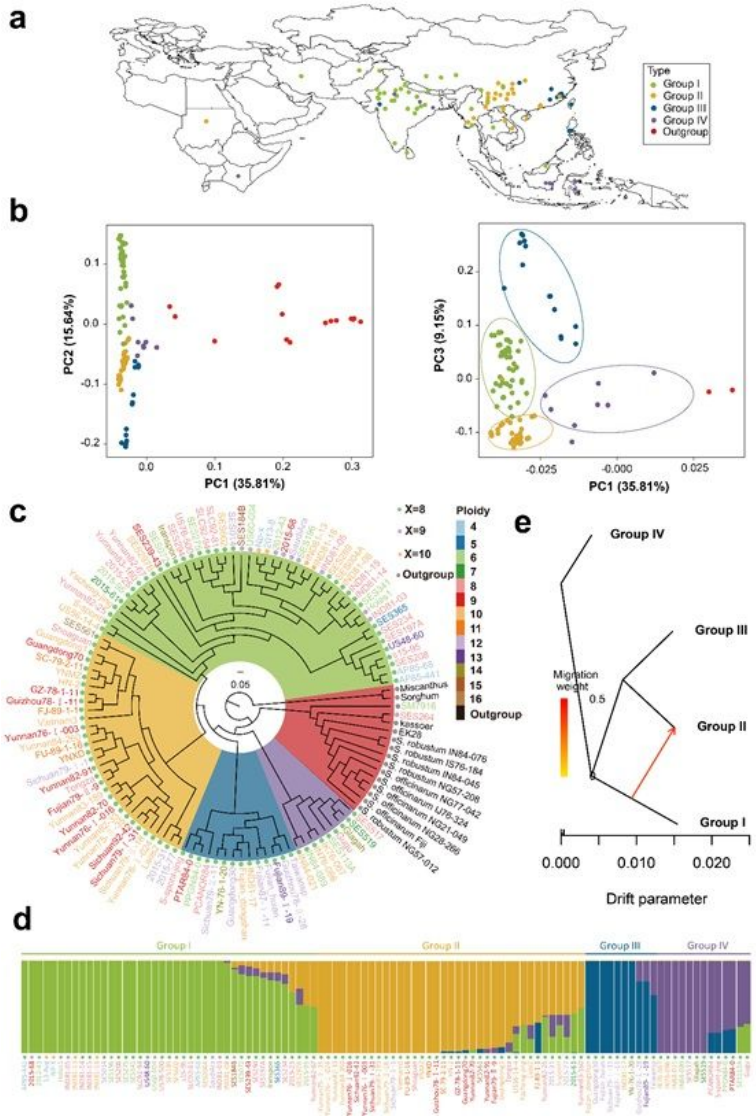

Figure 5

Genetic relationships among S. spontaneum population. (a) Geographic distribution of resequenced accessions. The dots of different colors represent the grouping results. (b) Principal component analysis of S. spontaneum accessions and the outgroup. PC1, PC2, PC3, the top three principal component with largest variance proportion explained. Dot colors are the same as in (a). (c) Phylogenetic tree based on whole-genome SNPs, with the colors of branch regions reflecting different groups (Group I, Group II, Group III, and Group IV). Colored dots in the outer ring indicate different ploidy levels while the gray dots represent the outgroup accessions. The basic chromosome numbers of x = 8, x = 9, or x = 10 for each accession are indicated as green, purple, or orange circles, respectively. Accessions with the same ploidy are represented in same color marked within the accession name. (d) Population structure analysis of representative S. spontaneum accessions and the outgroup for K = 4. (e) TreeMix analysis of the gene flow among the four S. spontaneum groups (Group I, Group II, Group III, and Group IV). The arrow corresponds to the direction of gene flow. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

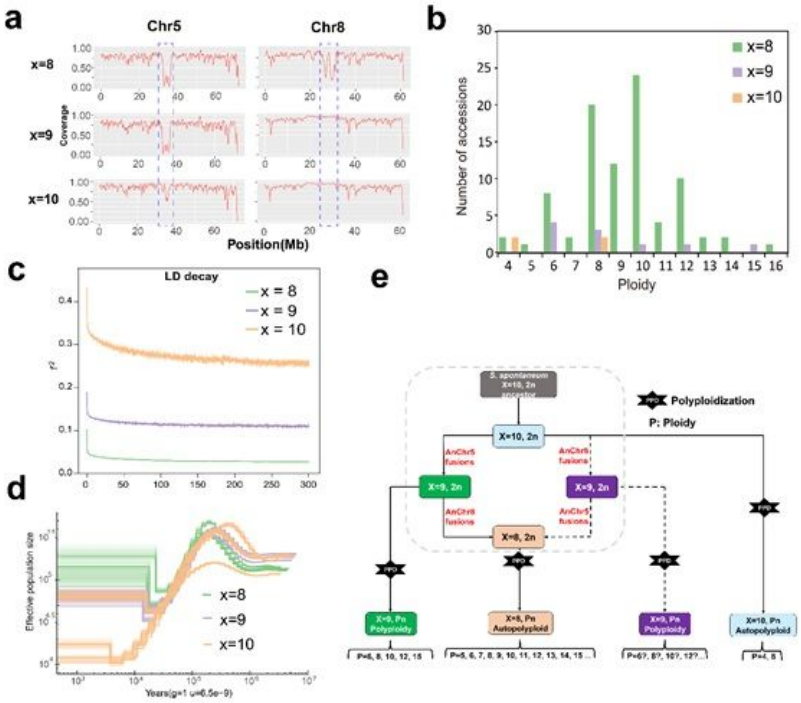

Figure 6

The population evolution in S. spontaneum. (a) Determination of basic chromosome number using read mapping coverage of re-sequenced S. spontaneum accessions. The regions marked with blue dotted boxes are diverged in S. spontaneum with basic chromosome numbers x = 8, x = 9, and x = 10. (b) The genetic background of 102 S. spontaneum accessions used for resequencing in the present study. The x-axis and the y-axis represent ploidy level and number of accessions, respectively. The columns in black, green, or red indicate the S. spontaneum accessions with basic chromosome numbers of x=8, x=9, or x=10, respectively. (c) Analysis of linkage disequilibrium (LD) decay in S. spontaneum populations with basic chromosome numbers of x = 8, x = 9, or x = 10. (d) MSMC-derived demographic history of S. spontaneum for 12 individuals: four accessions each are from a population with a basic chromosome number of x = 8, x = 9, or x = 10, respectively. (e) Schematic of the evolution of different basic chromosome numbers in S. spontaneum.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ExtendedData1GenefunctionannotationforS.spontaneumNpXgenome.xlsx
- ExtendedData2ThealleletableofS.spontaneumNpXgenome.xlsx
- ExtendedData4ThehomologgenescomparsionbetweenS.spontaneumNpXS.spontaneumAP85441andSorghum.xlsx
- ExtendedData3ThecomparsionofphytosynthesisrelatedgenefamiliesinS.spontaneumNpXS.spontaneumAP85441MiscanthusandSorghum.xlsx
- ExtendedData5Thetargetgenesandcorrespondfunctionannotationinselectivesweepregion.xlsx
- 03.SupplementaryFiguresXTables20210425.pdf