Chromosome-scale assembly and analysis of biomass crop *Miscanthus lutarioriparius* genome

Jiashun Miao

CAS Center of Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology,
Chinese Academy of Sciences

Qi Feng

National Centre for Gene Research <https://orcid.org/0000-0001-5450-1343>

Yan Li

Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences

Qiang Zhao

National Center for Gene Research

Congcong Zhou

Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences

Hengyun Lu

Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences

Danlin Fan

National Center for Gene Research & Institute of Plant Physiology and Ecology, Shanghai Institutes of
Biological Sciences, Chinese Academy of Sciences

Yiqi Lu

Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences

Qilin Tian

National Centre for Gene Research

Wenjun Li

Chinese Academy of Sciences

Qijun Weng

National Center for Gene Research

Lei Zhang

National Centre for Gene Research

Yan Zhao

Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences

Tao Huang

National Center for Gene Research & Institute of Plant Physiology and Ecology, Shanghai Institutes of
Biological Sciences, Chinese Academy of Sciences

Juan Yan

Key Laboratory of Plant Germplasm Enhancement and Speciality Agriculture, Wuhan Botanical Garden,
Chinese Academy of Sciences

Laigeng Li

Chinese Academy of Sciences <https://orcid.org/0000-0001-6924-4431>

Xuehui Huang

College of Life Sciences, Shanghai Normal University <https://orcid.org/0000-0003-0529-1532>

Tao Sang

Institute of Botany, Chinese Academy of Sciences, Beijing, China

Bin Han (✉ bhan@ncgr.ac.cn)

CAS Center of Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology,
Chinese Academy of Sciences <https://orcid.org/0000-0001-8695-0274>

Article

Keywords: chromosome-scale assembly, *Miscanthus lutarioriparius*, Oxford Nanopore sequencing, Hi-C technologies

Posted Date: September 30th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-81936/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Nature Communications on April 28th, 2021. See the published version at <https://doi.org/10.1038/s41467-021-22738-4>.

Chromosome-scale assembly and analysis of biomass crop *Miscanthus* *lutarioriparius* genome

Jiashun Miao^{12†}, Qi Feng^{1†}, Yan Li¹, Qiang Zhao¹, Congcong Zhou¹, Hengyun Lu¹, Danlin Fan¹,
Yiqi Lu¹, Qilin Tian¹, Wenjun Li¹, Qijun Weng¹, Lei Zhang¹, Yan Zhao¹, Tao Huang¹, Juan Yan³,
Laigeng Li¹, Xuehui Huang⁴, Tao Sang^{5*}, Bin Han^{1*}

¹ National Center for Gene Research, State Key Laboratory of Plant Molecular Genetics, CAS
Center of Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology,
Chinese Academy of Sciences, Shanghai 200233, China.

² University of Chinese Academy of Sciences, Beijing 100049, China.

³ CAS Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan
Botanical Garden, Chinese Academy of Sciences, Wuhan, Hubei 430074, China.

⁴ College of Life Sciences, Shanghai Normal University, Shanghai 200234, China.

⁵ Key Laboratory of Plant Resources and Beijing Botanical Garden, Institute of Botany, Chinese
Academy of Sciences, Beijing 100093, China.

[†] These authors contributed equally to this work.

* Correspondence should be addressed to B.H. (email: bhan@ncgr.ac.cn) or T.S. (email:
sang@ibcas.ac.cn).

Abstract

The genus *Miscanthus*, a rhizomatous perennial plant, has great potential for bio-energy production for its high biomass yield and strong stress resistance. Here, we generated the first chromosome-scale assembly of *Miscanthus lutarioriparius* Nandi by combining Oxford Nanopore sequencing and Hi-C technologies. The 2.07-Gb assembly covers 96.64% of the genome, with contig N50 of 1.71 Mb. By utilizing Hi-C data, 1,956.46 Mb (94.30%) of the assembled sequences were anchored and oriented to 19 pseudochromosomes of haploid *M. lutarioriparius*. The diploid *M. lutarioriparius* was identified to have a typical tetraploid genome structure. With the chromosome-level genome assembly, a series of structural variation, including chromosomal fragmentation, fusion and rearrangement relative to sorghum genome were clearly characterized. The tandem duplicate genes of *M. lutarioriparius* were enriched not only in the GO terms related to biotic and abiotic stress response, but also in the cell wall biosynthesis. Comparative genome analysis revealed that the gene families related to disease resistance, cell wall biosynthesis and metal ion transport were greatly expanded and evolved. The gene families associated with the distinctive features of *M. lutarioriparius* were analyzed in detail, involving cell wall biosynthesis, disease resistance and C₄ photosynthesis pathways. Besides the recent whole genome duplication, tandem duplication has played critical role in the expansion of size of these gene families investigated. While the transcriptome data showed that most duplicated gene members had extremely low level of expression in all transcriptome samples or the expected sample. The expansion of these families probably be an important genomic basis for the enhancement of those remarkable traits of *M. lutarioriparius*. These candidate gene members also provide valuable gene resource for function

analysis and genetic improvement for *M. lutarioriparius*.

Introduction

The genus *Miscanthus* contains approximately 20 species, among which 7 species are widely distributed in China^{1–3}. It is a rhizomatous perennial plant that has great potential for bio-energy production for its high biomass yield and strong stress resistance^{3,4}. It is one of a few C₄ photosynthetic plants that adapt to the chilling even cold conditions, which made *Miscanthus* a valuable genetic resource for sugarcane cold tolerance breeding^{3,5,6}. *Miscanthus* has been shown to be tolerant to heavy metals, which makes it to be an excellent material for phytoremediation of heavy metal contaminated in soil^{7–13}. *Miscanthus* is much more suitable for phytostabilization of heavy metal in soil, since the concentration of heavy metals in roots is higher than that in shoots^{14,15}.

Miscanthus lutarioriparius Nandi, an endemic *Miscanthus* species in middle and low ranges of the Yangzi River of China, can grow up to 7-meter tall and has the highest biomass production among four major *Miscanthus* species that widely distributed in China^{16–18}. It is found to have good paper-making properties, which is one way to change the shortage of raw materials for paper-making industry in China. The species is considered to be a promising second-generation energy crop because of its high biomass production and strong adaptability through high photosynthetic rates, water use efficiency and strong tolerance to drought and salt when growing on marginal land^{19–21}.

Miscanthus is self-incompatible (SI), which may contribute to the extensive environmental adaptation and high biomass production of *Miscanthus*^{3,22}. However, the heterozygosity in *Miscanthus* genome hindered the advances of the genome sequencing and assembly in the past. The lack of genomic resources limited our understanding of the genomic basis of distinctive traits, genome evolution and the process of molecular breeding in *Miscanthus*. In addition to the heterozygosity, there are other interfering factors in the genome assembly of *Miscanthus*. The ploidy and genome size of species in genus *Miscanthus* are significantly different^{23–25}. For instance, *M. sacchariflorus* could be with diploid or tetraploid^{24,26}. As for the species we study, *M. lutarioriparius* (2n = 2x = 38) possesses a relatively large genome size with abundant repetitive sequences, which made it difficult to achieve a high-quality genome assembly using short-read sequencing technologies. Recently, the single-molecular, real-time sequencing technologies and high-throughput chromosome conformation capture (Hi-C) technology have been used in combination to make significant advances in improving the assembly of large and complicated genomes at chromosomal level, especially for plant genomes²⁷.

Here we generated the first chromosome-level reference genome of *M. lutarioriparius* by combining Oxford Nanopore sequencing and Hi-C technologies. This chromosome-level genome assembly of *M. lutarioriparius* together with the assembly of *M. sinensis* (<https://phytozome.jgi.doe.gov/>) will facilitate the better utilization of *Miscanthus* genetic resource in multiple aspects.

Results

High quality genome assembly of *Miscanthus lutarioriparius*

Before the genome *de novo* assembly, genome survey based on k-mer frequency revealed a high level of repeat content (~67.30%) in *M. lutarioriparius* (Supplementary Table 1 and

Supplementary Fig. 1). The genome size estimated by k-mer statistics is about 2.19 Gb, close to 2.147 Gb determined using flow cytometry¹⁸.

The genome of *M. lutarioriparius* was *de novo* assembled using the long reads generated by PromethION platform of Oxford Nanopore Technologies and incorporated with Hi-C technology for scaffolding. We generated 307.71 Gb Nanopore long reads with a read N50 length of 32.2 kb (Supplementary Table 2,3 and Supplementary Fig. 2). For draft assembly improvement, three Illumina paired-end libraries with different insert-size were constructed, and a total of 205.74 Gb data were generated using Illumina HiSeq platform (Supplementary Table 4). The total length of contigs assembled by SMARTdenovo software is about 2,250.39 Mb with contig N50 size of 1.71 Mb, which is a little larger than the estimated genome size. Sequence comparison of the assembled initial contigs to three BAC sequences determined by Sanger sequencing showed good agreement with high sequence identity, ranging from 99.52% to 99.68% (Supplementary Fig. 3). Redundant sequences resulting from genome heterozygosity were collapsed during the following Hi-C anchoring process, where the total length of assembly reduced from 2,250.39 Mb to 2,074.63 Mb (without gaps and Ns). The final genome assembly consists of 919 scaffolds, which spans 2,074.80 Mb in total, with the scaffold N50 size of 113.46 Mb (Table 1). The final assembly covers 96.64% of 2,147 Mb genome on the basis of the flow cytometry¹⁸. About 94.30% of total sequences were anchored into 19 pseudochromosomes with sizes range from 61.78 Mb to 150.81 Mb (Fig. 1a and Supplementary Table 5). Syntenic blocks account for 87.36% of total coverage across the whole genome assembly (Fig. 1b and Supplementary Table 6).

To evaluate the reliability of the Hi-C scaffolding, we employed two Hi-C scaffolding programs LACHESIS²⁸ and 3D *de novo* assembly (3d-dna) pipeline²⁹ to anchor the contigs, respectively. The scaffolds constructed by two methods were massively consistent in contig order and orientation. LACHESIS and 3d-dna pipeline anchored and oriented 1,956.46 Mb (94.30%) and 1,806.79 Mb (93.93%) contigs into 19 pseudochromosomes, respectively (Fig. 1a and Supplementary Fig. 4). Except that Chr15 and Chr19 have relatively lower mean sequence identity (~85%), the rest scaffolds show very high sequence identity between two methods, providing confidence in the scaffolding results. Several small inversions were observed at the end of chromosomes Chr09, Chr13 and Chr15 when comparing two results (Supplementary Fig. 5). The LACHESIS assembly reached a LTR Assembly Index (LAI) of 12.11, which is better than that of 3d-dna pipeline (9.06), indicating a higher genome contiguity and sequence quality³⁰. The LACHESIS assembly was selected as the final genome assembly, since it is better than 3d-dna assembly by further quality evaluation (Supplementary Table 7).

We further verified the quality of the Hi-C scaffolding by analyzing the genomic synteny between *M. lutarioriparius* Hi-C scaffolding assembly and sorghum genome. The results showed a conserved synteny between *M. lutarioriparius* and sorghum in both coding sequences (CDS) and genomic sequence scale (Fig. 2a and Supplementary Fig. 6). The LAI score of assembly constructed by Hi-C scaffolding was 12.11 (Supplementary Fig. 7), reaching to the criterion of reference quality (with LAI score ranges from 10 to 20)³⁰. A total of 1,339 (97.4 %) complete gene models among 1,375 conserved genes in Benchmarking Universal Single Copy Orthologs (BUSCO) were recalled in our final genome assembly (Table 1). Furthermore, three Illumina short reads libraries with different insert-size were mapped back to the assembly, and overall alignment rate were 99.82%, 99.79% and 99.78%, respectively. RNA sequencing data generated for genome annotation was also used to assess the quality of assembly by mapping back to the assembly, and overall alignment rates

ranged from 89.24% to 95.37% in nine libraries (Supplementary Table 8). Given the above, the genome assembly of *M. lutarioriparius* presents high contiguity and sequence quality.

The average GC content of *M. lutarioriparius* genome is about 45.46% (Supplementary Table 9). Relatively lower GC content were observed for chromosome 9 and 10 (Supplementary Fig. 8). The potential centromeric sequences of *M. lutarioriparius* were assembled for all 19 chromosomes (Supplementary Fig. 9 and Supplementary Table 10). The centromeric satellite sequences of *M. lutarioriparius* is a highly repetitive sequence of 137 bp unit and the sizes vary from 106.8 kb (chromosome 19) to 7.7 Mb (chromosome 10). Telomere sequence of chromosome 10 was assembled in our assembly. A total of 1,917 *Arabidopsis*-type 7-bases TTTAGGG³¹ repeats were detected at one end of chromosome 10 (Supplementary Fig. 9).

Table 1. Statistics of genome assembly.

	Hi-C scaffolds
Assembly features	
Number of scaffolds	919
Total size of scaffolds	2074.80 Mb
Longest scaffold	150.81 Mb
Shortest scaffold	16.94 kb
Mean scaffold size	2.26 Mb
Median scaffold size	130.09 kb
N50 scaffold length	113.46 Mb
L50 scaffold count	8
Scaffold GC content	45.46%
Scaffold N content	0.01%
Percentage of assembly in scaffolded contigs	94.30%
Average number of contigs per scaffold	2.9
BUSCO (complete)	97.40%
LTR Assembly Index (LAI)	12.11
Gene models	
Number of gene models	68,328
Mean coding sequence length	1,215 bp
Mean number of exons per gene	4.77
Mean exon length	255 bp
Mean intron length	727 bp
Non-protein coding genes	
Number of miRNA gene	521
Number of tRNA gene	1,164
Number of rRNA gene	257
Number of snoRNA gene	970
Number of snRNA gene	97

A total of 68,328 gene models were predicted, with an average coding sequence (CDS) length of 1,215 bp, of which about 94.75% (64,742) were assigned to 19 chromosomes (Supplementary Table 11). The length distribution of CDS, exons and introns of *M. lutarioriparius* are similar to that of *S. spontaneum* and *S. bicolor* (Supplementary Fig. 10), providing confidence in the gene prediction results. The average GC content of the coding sequence (CDS) of *M. lutarioriparius* (56.40%) is similar to that of *S. spontaneum* (56.40%), *S. bicolor* (56.74%) and *S. italica* (56.50%), but higher than that of *A. thaliana* (44.45%) (Supplementary Table 12). The GC content and GC3s (GC levels of third codon positions) of CDS showed a bimodal distribution in *M. lutarioriparius* (Supplementary Fig. 11, 12).

Previous studies revealed that the various GC content of genome region may have different biological properties, such as gene density, composition of repeat sequences and recombination³²⁻³⁴. Strong positive linear correlation (Pearson correlation coefficient = 0.95, FDR = 8.25e-10) was detected between GC content and gene density in *M. lutarioriparius* genome (Supplementary Fig. 13). Notably, gene density of chromosome 9 (25.19 genes / Mb) and 10 (26.27 genes / Mb) are significantly lower than the average gene density of whole genome (32.93 genes / Mb).

We evaluated the completeness of gene prediction with 1,375 BUSCO genes from embryophyta_odb10, of which 1,350 BUSCOs (98.2%) were complete (Supplementary Table 13). Among 68,328 predicted proteins, 63,076 (92.31%) were classified by InterProScan³⁵, 63,590 (93.07%) were classified by eggNOG-mapper³⁶, 22,902 (33.52%) were functional annotated in KEGG database and 39,173 (57.33%) were assigned with GO annotation. A total of 4,031 transcript factors were annotated. In addition, 1,164 tRNA genes, 257 rRNA genes, 521 microRNA (miRNA) genes, 98 small nuclear RNAs (snRNAs) genes and 970 small nucleolar RNAs (snoRNAs) genes were predicted by integrating different methods (Table 1).

A total of 1.34 Gb (64.39% of nucleus genome) interspersed repeats were identified in *M. lutarioriparius* genome (Supplementary Table 14). The long-terminal repeat retrotransposons (LTR-RTs) are the most abundant type of repetitive sequences in *M. lutarioriparius*, spanning 46.78% of the nucleus genome. A total of 8,848 intact LTR-RTs were classified in our assembly, and the frequency distribution of intact LTR-RTs insertion time showed a burst of LTR-RTs 1-2 million years ago (MYA) (Supplementary Fig. 14). The largest LTR-RTs superfamily *Gypsy*, comprising ~35.2% of the genome, is concentrated near the putative centromeres (Supplementary Fig. 15d). The second largest superfamily of LTR-RTs *Copia* comprises ~11.6% of the genome (Supplementary Fig. 15c). Other interspersed repeats, LINEs (Long Interspersed Nuclear Elements) and SINEs (Short Interspersed Nuclear Elements) occupy 1.21% and 0.16% of the genome, respectively (Supplementary Fig. 15a, b). The sequence divergence rate of LINEs is relatively higher (median = 20.6%) compared to other types of transposable elements (Supplementary Fig. 16). DNA transposons comprises ~9.64% of the *M. lutarioriparius* genome. Strikingly, DNA transposons show obvious enrichment on the arms of the GC-poor chromosome 9 and 10 (Fig. 1b and Supplementary Fig. 15e, f). Whereas there is no significant correlation between the DNA transposons content and GC content for both chromosome 9, 10. A total of 19,062 miniature inverted-repeat transposable elements (MITEs) was identified, spanning 0.23% of the genome (Supplementary Table 15 and Supplementary Fig. 17). In total, 517,973 tandem repeats were identified, spanning 3.74% of the genome. (Supplementary Table 16 and Supplementary Fig. 9).

Genomic structure variation in *Miscanthus lutarioriparius*

Investigation of collinear orthologs between *M. lutarioriparius* and sorghum confirmed the occurrence of the recent whole genome duplication (WGD) event in *M. lutarioriparius* (Fig. 2a). Except that chromosome 7 of *M. lutarioriparius* has syntenic relationship with two sorghum chromosomes (SbChr04 and SbChr07), the rest chromosomes are aligned to one sorghum chromosome (Fig. 2a). There are 65,044 (95.19%) *M. lutarioriparius* genes having at least one syntenic gene in *S. biocolor*. About 87.05% (29,710) of sorghum genes have 2 syntenic genes in *M. lutarioriparius* genome, showing a predominant 2-to-1 pattern of syntenic depth between *M. lutarioriparius* and Sorghum, and revealing the tetraploid nature of the diploid *M. lutarioriparius* (Fig. 2a). 2,802 (8.21%) sorghum genes have more than two syntenic genes in *M. lutarioriparius* genome, which presumably resulted from the segmental duplication, tandem duplication or single-gene duplication occurred in *M. lutarioriparius* genome after it split with sorghum. Only about 0.68% (233) of genes in sorghum have no syntenic gene in *M. lutarioriparius*, indicating that few genes were lost in *M. lutarioriparius* after it split with sorghum.

From the genome comparison between *M. lutarioriparius* and *S. biocolor*, one chromosome fusion event occurred after the recent WGD and several intra-chromosome rearrangements relative to sorghum were identified in *M. lutarioriparius* (Fig. 2a and Supplementary Fig. 6). Two large inversions were observed on the two homologous chromosomes of *M. lutarioriparius* (MlChr07 and MlChr08), which is consistent with the research carried out in *Miscanthus sinensis* based on the RNA-Seq data³⁷ (indicated with orange circles in Fig. 2a). The common ancestor's chromosomes of SbChr04 and SbChr07 were fused into one chromosome MlChr07 in *M. lutarioriparius*, which probably occurred after the *M. lutarioriparius* recent WGD, resulting in MlChr07 aligning to two sorghum chromosomes (Fig. 2a). The preserved MlChr08 and MlChr13 are intact homologues of SbChr04 and SbChr07, respectively (Fig. 2a). Besides, inter-chromosomal exchanges between MlChr09/10 and MlChr14/15 were observed, which probably occurred prior to the *M. lutarioriparius* recent WGD (indicated with green circles in Fig. 2a).

To dating the recent WGD in *M. lutarioriparius*, genomic synteny analysis based on the self-comparison of *M. lutarioriparius* protein-coding genes was carried out. The two peaks of the synonymous substitution values (K_s) frequency distribution indicate that an ancient burst gene duplication occurred in *M. lutarioriparius* prior to the recent WGD (Fig. 2b). The calculation of K_s for *M. lutarioriparius* versus *M. lutarioriparius* collinear gene pairs suggested that the recent WGD event may date back to ~6.15 MYA (Fig. 2b).

We further analyzed the duplicate gene origins in *M. lutarioriparius*. The results suggest that the WGD/segmental duplication is the predominant duplicate gene origin type (63.96%, 43,704) compared to the other three types: dispersed duplication (18.59%, 12,700), tandem duplication (6.39%, 4,365) and proximal duplication (6.24%, 4,267) (Supplementary Fig. 18a). Notably, *Miscanthus* has the highest proportion of tandem duplication origin compared to the other eight investigated taxa (Supplementary Fig. 18b). The chromosome 9 and 10 have the highest proportion of proximal and tandem duplication compared to that of other chromosomes (Fig. 2c). The analysis of GC3s and GC content of genes of different duplication origins indicates that the tandem duplicates have the highest GC3s value and lowest effective number of codons (ENC/Nc) compared to other duplication types (Supplementary Fig. 19).

To find out if genes of different duplication origins have biological function preference, Gene

Ontology (GO) enrichment analysis were carried out. GO analysis suggested that genes created through WGD/segmental duplication were enriched with the (GO) terms such as “cell redox homeostasis”, “intracellular protein transport”, “signal transduction”, “metal ion transport” and “photosynthesis” (Supplementary Fig. 20). Previous study has demonstrated that the tandem arrayed genes in both rice and *Arabidopsis* genomes were enriched in the function of “biotic and abiotic stress”³⁸. While gene duplicates originated from tandem duplication in *M. lutarioriparius* were enriched not only in the GO terms related to biotic and abiotic stress response, but the cell wall biosynthesis (Fig. 2d, Supplementary Fig. 21, 22). Tandemly arrayed genes are thought to be volatile after gene duplication, so the retained tandem genes may indicate functional importance³⁹. These significantly enriched GO terms of tandem gene duplicates were closely associated with the distinctive features of *M. lutarioriparius*: higher environmental adaptability, disease resistance and lignocellulosic biomass. It is noteworthy that a GO term named “recognition of pollen”, also known as “self-incompatibility”, was significantly enriched in tandem duplicates, indicating the importance of self-incompatibility in *M. lutarioriparius* (Fig. 2d). The duplicate genes resulted from proximal duplication were mainly enriched in the function of cell wall biosynthesis, such as “transferring hexosyl groups” and “polysaccharide binding” (Supplementary Fig. 23). These gene duplicates classified as dispersed were enriched in the functions such as “DNA repair”, “telomere maintenance”, “nucleosome assembly”, “protein heterodimerization activity”, “DNA helicase” (Supplementary Fig. 24).

Genome evolutionary history of *Miscanthus lutarioriparius*

To investigate the genome evolutionary history of *M. lutarioriparius*, gene family clustering was carried out using *M. lutarioriparius* and seven other representative angiosperm species, including an outgroup species (*Arabidopsis thaliana*). From the gene family clustering result, 122 single-copy orthologs shared by *M. lutarioriparius* and seven other green plants species were used for phylogenetic reconstruction and species divergence time estimation, which shows *M. lutarioriparius* is closest relative to *S. spontaneum*⁴⁰. We estimated that *M. lutarioriparius* and *S. spontaneum* shared a common ancestor ~7.97 MYA. The divergence of *M. lutarioriparius* and *S. bicolor* was estimated to occur ~9.59 MYA (Fig. 3b).

Comparative genomics of gene family in *Miscanthus lutarioriparius*

To reveal the genomic basis of *M. lutarioriparius* distinctive phenotypes, we investigated the evolution of gene families by characterizing unique and shared genes families among different plants. Through comparing gene families of seven grass subfamilies, including Pooideae (*Brachypodium distachyon*), Ehrhartoideae (*Oryza sativa*), Panicoideae (*Zea mays*, *Sorghum bicolor*, *Setaria italica*, *Saccharum spontaneum* and *Miscanthus lutarioriparius*) and one dicot plant (*Arabidopsis thaliana*), we characterized 57,710 genes of *M. lutarioriparius* into 21,515 gene families (Fig. 3a). Further comparison of the species in subfamily Panicoideae, including *M. lutarioriparius*, *Z. mays*, *S. bicolor*, *S. italica* and *S. spontaneum*, revealed 16,080 gene families distributed among all five genomes and 144 gene families that were unique to *M. lutarioriparius* (Fig. 3a). Notably, ten of these 144 gene families contain NB-ARC domain, a signaling motif shared by most plant resistance (*R*) gene productions^{41–44}, implying stronger disease resistance might have

evolved in *M. lutarioriparius*. And three of these 144 gene families were annotated to be cytochrome P450, a multifunctional oxidase which is well known for its important role in the biosynthesis of defensive compounds, hormones, and fatty acids in plants^{45,46}. Further GO enrichment analysis of these 144 gene families revealed totally six significant GO terms (adjust *P* value < 0.05), including “peroxidase activity”, “response to oxidative stress”, “ubiquitin-like modifier activating enzyme activity”, “cellular amino acid metabolic process” and “DNA integration and carboxy-lyase activity” (Fig. 3c). Peroxidases are demonstrated to be involved in plant response to biotic and biotic stresses^{47–49}.

Further gene family expansion and contraction analysis for *M. lutarioriparius* was carried out. In total, 9,509 gene families in *M. lutarioriparius* were identified as expanded, whereas 3,228 gene families were identified as contracted (Fig. 3b), of which 219 were identified as rapidly evolving families (211 expanded and 8 contracted). These rapidly evolving gene families may give us some hints about the harsh environmental adaptability of *M. lutarioriparius*. Function annotation of these 211 rapidly expanded gene families revealed that they are involved in a variety of plant biological processes, but most notably in dealing with diverse biotic and abiotic stresses, particularly disease resistance (Fig. 3d). Among these 211 rapidly expanded gene families, 21 gene families composed of 334 genes contained the NB-ARC domain. Four rapidly expanded gene families composed of 56 genes were annotated to be related to the xylanase inhibitor, which could be induced by biotic and abiotic stimulus, and involve in plant defense against fungal pathogens⁵⁰. Other stress resistance related function annotation for these rapidly evolving gene families included “salt stress response/antifungal”, “wound induced protein”, “ubiquitin family”, “WRKY DNA-binding domain”, “Cytochrome P450”, “thaumatin family”, and “terpene synthase” (Fig. 3d).

The GO enrichment analysis for the 9,509 expanded gene families revealed that the significantly enriched GO term with the largest gene members (155) in biological process category was “metal ion transport” (Fig. 3e). Notably, GO terms related to the cell wall biosynthesis were present in the enrichment results of both molecular function and cellular component category (Fig. 3e and Supplementary Fig. 25).

Most of disease resistance genes cloned in plants are found to encode nucleotide-binding-site-leucine-rich-repeat (NBS-LRR) proteins⁵¹. In *M. lutarioriparius* genome, 547 NBS-LRR coding genes were identified, which is more than that of rice (~500)⁵², sorghum (211-346)^{53–55}, maize (137)⁵³ and *S. spontaneum* (361)⁴⁰ (Supplementary Table 17). *M. lutarioriparius* NBS-LRR genes mostly encode the CC-NBS-LRR (262). A total of 234 (42.8%) NBS-LRR genes are enriched on chromosome 9 and 10 (homologues of sorghum chromosome 5), most of which fall into multi-gene clusters, supporting the conservation of R gene location⁵⁴ (Supplementary Fig. 26c). About 17.4% (95) NBS genes were classified into tandem duplicates, which was significantly higher than that of genome-wide level (6.39%), indicating the important role of tandem duplication in the expansion of NBS genes in *M. lutarioriparius*. Phylogenetic analysis based on NBS domain showed no distinct clades, and yielded a “star-like” topology (Supplementary Fig. 26a,b,d,e), which was similar to that of rice and sorghum⁵³.

Complex carbohydrates of plants are the important food sources and renewable biofuel feedstock. Carbohydrate active enzymes (CAZymes) play critical role in complex carbohydrate metabolism^{56,57}. A total of 2,919 (4.3%) genes encoding CAZyme domains were identified in *M. lutarioriparius*, which was the most abundant among the 12 species investigated (Supplementary Fig 27). GTs comprise approximately 46% of the CAZyme domain content in *M. lutarioriparius*,

with GHs having a relative frequency of 32%, CEs have a smaller proportion (4.6%) compared to other species. A total of 47 GT families are present in *M. lutarioriparius*, among which twelve GT families (GT1, GT2, GT47, GT31, GT8, GT0, GT61, GT106, GT4, GT77, GT37 and GT34) account for more than 80% of all GT members. GT1 is the largest GT family with 396 (27.54%) members in *M. lutarioriparius*, and GT1 subfamily also constitute the majority of GT genes in *Arabidopsis* (> 25%) and rice (> 35%)⁵⁸ (Supplementary Fig. 28). In addition, 27 genes coding cellulose synthases (CesA) were identified in *M. lutarioriparius* (Supplementary Table 18). Transcriptome analysis showed almost half CesA genes have specific high expression in the middle part of internode (Supplementary Fig. 29b), implying the cellulose synthesis may be more active in the middle part of internode. A total of 90 genes encoding cellulose synthase-like (Csl) were identified in *M. lutarioriparius*, which was larger than that of maize (33), rice (35) and *Arabidopsis* (29) (Supplementary Table 19). Phylogenetic analysis separated all *M. lutarioriparius* Csl genes into six distinct groups and revealed that CslD group is most closely related to CesA family among all Csl groups, but CslD and CesA differed greatly in gene structure (Supplementary Fig. 30). Tandem duplication, WGD/segmental duplication and proximal duplication greatly expanded the gene families of Csl and CesA in *M. lutarioriparius* (Supplementary Fig. 29, 30 and 31). So, we speculated the higher cellulose content of *M. lutarioriparius* may be contributed by the expansion of the cellulose synthase related gene families (Supplementary Table 20).

Lignin, a component of plant cell walls, is an amorphous polymer that responsible for maintaining the rigidity and mechanical structure of cell walls, providing a barrier against pathogens. Lignin can directly reduce the entry of heavy metals into the root system, enhancing the tolerance of plants to heavy metals⁵⁹. In *M. lutarioriparius*, 333 genes belonging to ten gene families were inferred to involve in lignin biosynthesis, which is greater than that of sorghum (141) and rice (155)⁶⁰ (Supplementary Table 21). Various types of gene families are distributed closely on chromosomes, which may be benefit to the improvement of the lignin biosynthesis efficiency (Supplementary Fig. 32a).

C₄ pathway genes of *M. lutarioriparius*

Genes involved in the C₄ photosynthesis pathway of *M. lutarioriparius* were identified on the basis of homology to C₃ and C₄ pathway orthologs in rice and sorghum. The *M. lutarioriparius* genome contains 55 putative genes involved in nicotinamide adenine dinucleotide phosphate-malic enzyme (NADP-ME) C₄ pathway carbon fixation, which included ten carbonic anhydrase (CA) encoding genes, twelve phosphoenolpyruvate carboxylase (PEPC) encoding genes, six phosphoenolpyruvate carboxylase kinase (PPCK) encoding genes, three pyruvate orthophosphate dikinase (PPDK) encoding genes, six pyruvate phosphate dikinase regulatory protein (PPDK-RP) encoding genes, thirteen NADP-linked malic enzymes (NADP-ME), three NADP-linked malate dehydrogenase (NADP-MDH) and two RbcS (Supplementary Notes and Supplementary Table 22).

Most C₄ photosynthesis genes in *M. lutarioriparius* were mainly expanded through the WGD, leading to the number of them was almost twice that of sorghum. Proximal duplication, tandem duplication and segmental duplication further expanded the C₄ photosynthesis gene families in *M. lutarioriparius* (Supplementary Note). Most non-C₄ isoforms showed extremely low expression level among all transcriptome samples compared to these C₄ isoform genes. The C₄ isoform gene duplicates created by recent WGD showed a similar high expression in leaf sample, implying both

gene duplicates probably play role in the C₄ photosynthesis, which might contribute to *Miscanthus* C₄ photosynthesis at low temperature. While some non-C₄ isoforms of PPCK family showed distinctive expression patterns (Supplementary Note). For instance, the non-C₄ isoform M112G018490 had the highest expression level in root sample. The variation of expression pattern may indicate the neofunctionalization followed the C₄ photosynthesis gene duplication.

Phylogenetic analysis based on chloroplast genome

A nearly complete chloroplast genome of *M. lutarioriparius* assembled using the Illumina short reads derived from whole genome sequencing data. The assembled chloroplast genome is 142,989 bp in length, with 123 protein-coding, 71 tRNA and eight rRNA genes annotated (Supplementary Fig. 33). Phylogenetic analysis by integrating other released chloroplast genomes in *Miscanthus* genus confirmed our sequencing sample belonging to *M. lutarioriparius* (Fig. 4, Supplementary Fig. 34 and 35, Supplementary Table 23). Phylogenetic analysis based on chloroplast genome is of great value in providing support for evolutionary relationships for species in *Miscanthus* genus. The accessions we investigated were divided into three distinct groups, one group containing the accessions from *M. sinensis* (Type I and II), *M. sacchariflorus* (Type I), *M. transmorrisonensis* and *M. floridulus*, one group containing the accessions from *M. lutarioriparius*, *M. sacchariflorus* (Type II) and *M. × giganteus* and another group containing the accessions from *M. capensis*, *M. junceus* and *S. spontaneum* (Fig. 4). The *M. sinensis* accessions were separated into two subgroups (Type I and II) distributed in two clades, suggesting higher genetic diversity in *M. sinensis*. All but one accessions of *M. sacchariflorus* were divided into two distant groups, one is closely related to the Type I *M. sinensis* accessions, while the other is closely related to the accessions from *M. × giganteus* and *M. lutarioriparius*. *M. × giganteus* was thought to be the interspecific hybrid of diploid *M. sinensis* and tetraploid *M. sacchariflorus*, and the maternal lineage was inferred to be the tetraploid *M. sacchariflorus*^{23,61}. So, we inferred these *M. sacchariflorus* accessions that closely related to *M. × giganteus* accessions were tetraploid. Our phylogeny analysis based on chloroplast genome sequences indirectly supports the view of point that tetraploid *M. sacchariflorus* and diploid *M. sacchariflorus* are taxonomically different, which has been proposed in year 2014²⁴. In our phylogenetic analysis, *M. lutarioriparius* and one accession from *M. sacchariflorus* (MH820376.1) are the closest relatives. Strikingly, *M. × giganteus* accessions are most closely related to *M. lutarioriparius* and one accession of *M. sacchariflorus* (MH820376.1). To better uncover the complicated relationship in genus *Miscanthus*, more accessions and geographical origin information still need to be included in the phylogenetic analysis.

Discussion

Here, we generated a qualified chromosome-scale assembly of the *M. lutarioriparius* genome using the long reads generated by Oxford Nanopore Technologies platform and Hi-C technology. The advance in genomic sequencing and assembly technologies makes it possible to produce chromosome-scale scaffolds in an accurate and cost-effective way. The chromosome-scale genome assembly of *M. lutarioriparius* provided crucial information for comparative genome studies for genus *Miscanthus*. We believe this study will contribute to the parsing of the genomic basis of distinctive traits owned by *M. lutarioriparius* and the development of genome-assisted molecular

breeding program for *Miscanthus*.

The recent whole genome duplication and tetraploid genetic structure of diploid *M. sinensis* have been clearly demonstrated by genetic linkage map generated by both genotyping-by-sequencing (GBS) method and RNA-Seq based makers^{37,62}, and the linkage group 8 of *M. sinensis* was revealed to align to two sorghum chromosomes (7 and 8), indicating a chromosomal fusion event occurred in *M. sinensis* after the WGD^{37,62}. Here, we uncovered that *M. lutarioriparius* owns very similar genome structure with *M. sinensis* at chromosome level, as well as sharing the same base chromosome number 19, indicating the split of these two species occurring in a very short time. The other most common *Miscanthus* species, such as *M. sacchriflorus*, *M. floridulus*, also share the base chromosome number 19². A parsimonious origin of *Miscanthus* is thought to be single origin, which was proposed by Swaminathan et al.³⁷.

We compared the sequence similarity between the two homoeologous chromosomes in *M. lutarioriparius*, relatively low values (31.66% to 33.47%) were present for all pairs (Supplementary Table 23). In addition, the sequence length of the homoeologous chromosomes is significantly different for most chromosome pairs, among which the pair Chr14 and Chr15 had the greatest difference (Supplementary Table 5). Swaminathan et al. thought that polyploidizations in the lineages of multiple *Miscanthus* species were generated by the hybridization of closely related species in *Miscanthus* in allopolyploid fashion, by which their original chromosomal pairing patterns in the larger genome can be retained³⁷. All up, we speculate that the high sequence divergence between the homoeologous chromosomes in *M. lutarioriparius* is much more likely to be created by the hybridization of two closely related species. We expected that with more and more genomic data of genus *Miscanthus* released by the scientific community, the species containing the two sub-genomes of *M. lutarioriparius* could be discovered. Notably, Himalayan accessions where 2N = 40 (Amalraj and Balasundaram 2006) might represent an intermediate genome situation of *Miscanthus* before the chromosome fusion event, which may contribute to the further understanding of the evolution of *Miscanthus* genome.

Phylogeny based on chloroplast genome sequence provides a new perspective for the interspecific relationship of genus *Miscanthus*. Interestingly, *M. × giganteus* is closely related to *M. lutarioriparius* based on the chloroplast genomes. More molecular evidences and samples are needed to determine the taxonomic status of *M. lutarioriparius* and other species in genus *Miscanthus*.

We believe that our genome data will be helpful to functional genomic study and genetic improvement for *M. lutarioriparius*.

Figures

Fig. 1 Overview of the *M. lutarioriparius* genome assembly.

(a) Hi-C based contig anchoring of *M. lutarioriparius* genome. Post-clustering heat map showing density of Hi-C interactions between contigs from LACHESIS pipeline. (b) Circos plot of *M. lutarioriparius* genome assembly. The outermost layer of colored blocks is a circular representation of the 19 pseudo-chromosomes, with scale mark labeling each 5 Mb. The LTR Assembly Index (LAI) score is shown in (A). GC content density (B), gene density (C), *Gypsy* density (D), *Copia* density (E), DNA TE density (F) are calculated using 100-kb non-overlap window. The innermost layer shows inter-chromosomal synteny, with colored upper-layer links representing syntenic blocks generated from the recent *M. lutarioriparius* WGD, and the gray low-layer links representing homologs generated from older duplications.

Fig. 2 Whole genome duplication events within *M. lutarioriparius* genome

Fig. 3 Comparative genomics of gene family of *M. lutarioriparius*

(a) Venn diagram of shared orthologous genes families among five Panicoideae species genomes.

(b) Inferred phylogenetic tree reconstructed using 122 single-copy ortholog genes shared by 8 species identified using OrthoFinder⁶³. Divergence timing were indicated using transparent blue bars at the internodes with 95% highest posterior density (HPD). The number of expanded and contract gene families was marked with plus and minus ahead the digitals respectively.

(c) GO enrichment results of 144 *M. lutarioriparius* specific gene families.

(d) Functional annotation for the rapidly evolving (expanded) gene families in *M. lutarioriparius*. Left panel shows the gene family size among 8 species and right panel shows the functional annotation for gene families. Considering that a gene family may contain multiple Pfam domain annotations, the Pfam domain annotation would be retained only if more than half of the members of *M. lutarioriparius* gene family have that function domain. These Pfam domain annotations with unknown function would be removed for its uninformative. Pfam domain annotations related to transposase were removed since it is hard reach conclusions concerning development or environment adaptation directly.

(e) GO enrichment results of 9,509 expanded gene families of *M. lutarioriparius* (biological process category).

Fig. 4 Time-calibrated phylogeny analysis based on the chloroplast genomes

The calibration point was indicated using red point (data source: <http://www.timetree.org/>). The light blue bars at nodes correspond to the 95% highest posterior density (HPD) of the age of the nodes. The bottom axis is in millions of years (Myr). The NCBI Reference IDs were indicated in the parentheses. Three background colors were used to indicate three distinct groups in genus *Miscanthus*. The star was used to indicate the chloroplast genome assemblies of our sequencing accession, which were reconstructed directly from the Illumina paired-end reads derived from whole genome sequencing using a baiting and iterative mapping approach, by which two reference were used (*M. junceus* and *S. spontaneum*).

Methods (Online methods)

Plant material

The *M. lutarioriparius* plant used in this study was sampled in Honghu city, Hubei province, China and then transplanted in National Center for Gene Research, CAS, Shanghai, China.

Oxford Nanopore sequencing libraries construction and sequencing

High molecular weight (HMW) genomic DNA (gDNA) was extracted from young leaf samples of *M. lutarioriparius* with CTAB extraction method⁶⁴. The gDNA was then purified with Qiagen Genomic-tip100/G (Cat ID: 10243, Qiagen, Germany) in accordance with the manufacturer's instruction. The DNA fragments with length range from 20 to 50 kb were selected using BluePippin Size-Selection System (Sage Science Inc., Beverly, MA, USA). The Nanopore libraries were then constructed using Oxford Nanopore LSK-109 kit following the manufacturer's instruction. Nanopore libraries were sequenced on PromethION platform with Flow cell R9.4.1 chemistry and guppy software. In total, 307.71 Gb raw Nanopore data was generated in this experiment. After

adapter trimming and quality filtering, a total of 280.84 Gb clean data were kept, which is equivalent to approximately $128 \times$ genome coverage (k-mer frequency estimated genome size: 2.19 Gb) (Supplementary Table 2,3).

Illumina libraries construction and sequencing

Genomic DNA was extracted from *M. lutarioriparius* young leaves using Hi-DNAsecure Plant Kit according to the manufacturer's instructions (Tiangen Biotech, Beijing, China). Three different insert-size Illumina DNA-seq libraries for draft assembly improvement (error correction) were constructed using the KAPA HyperPrep Kit (Kapa Biosystems, Roche, USA) following the manufacturer's recommendations. Two libraries were sequenced on an Illumina HiSeq 2500 platform under paired-end 250 bp mode. The other two libraries were sequenced under paired-end 150 bp mode. A total of 205.74 Gb raw data were generated. After quality control by using fastp⁶⁵ (version 0.12.6) software with default parameters, 172.52 Gb clean data were kept for draft assembly improvement.

To facilitate genome annotation, we performed mRNA sequencing of nine different *M. lutarioriparius* tissues, which includes leaf, spikelet, root, internode (divided into upper, middle and lower parts), rhizome, lateral bud and seedling. Up to 400 µg total RNA per sample was purified for each sample by using TRIzol-based method (Invitrogen, CA, USA) and then subjected to DNase I treatment. The quality of extracted RNA per sample was evaluated using 1% agarose gel. Before polyA mRNA enrichment, the integrity of RNA was checked using Agilent 2100 bioanalyzer (Agilent Technologies, Inc., CA, USA). The polyA mRNA per sample was firstly enriched following the protocol of NEBNext[®] Poly(A) mRNA Magnetic Isolation Module (New England Biolabs #7490S, MA, USA). RNA Sequencing libraries were then constructed using NEBNext[®] Ultra[™] II RNA Library Prep Kit for Illumina (New England Biolabs #E7775, MA, USA) following the manufacturer's illustration. The concentration of RNA libraries was quantified using the Qubit[®] 3.0 Fluorometer. Libraries were then sent for 150 bp paired-end sequencing on an Illumina HiSeq 2500 platform. In total, 95.12 Gb raw data were generated for nine tissue samples.

Hi-C library construction and sequencing

The same plant of *M. lutarioriparius* used for Oxford Nanopore sequencing was used for Hi-C libraries construction. Tender leaves were collected for Hi-C libraries construction with a modified Hi-C method^{66,67}. Briefly, young leaf sample was fixed with formaldehyde. Fixed sample was then lysed, and the cross-linked DNA was digested with restriction endonuclease DpnII (New England Biolabs, USA) overnight. Biotin-labeled bases were introduced during the sticky ends repairing process. The interacting DNA fragments were ligated to form chimeric junctions. DNA fragments were purified and physically sheared to a size of 300-700 bp. DNA fragments tagged with biotin were enriched with beads and then sent to DNA sequencing libraries construction. The libraries were sent to sequenced on an Illumina HiSeq 4000 platform under paired-end 150 bp mode. In total, 347.76 Gb ($\sim 159 \times$ genome coverage) clean data was generated after quality control by fastp⁶⁵ (version 0.12.6) with default parameters. The quality of Hi-C experiments was evaluated using HiC-Pro⁶⁸ (Supplementary Table 24,25).

Estimation of genome size and heterozygosity

The genome size of *M. lutarioriparius* was firstly estimated using k-mer frequency distribution generated from Illumina short reads. The program jellyfish⁶⁹ (version 2.2.9) was used to calculate 17-mer frequency distribution. GCE (Genome Characteristics Estimation) software⁷⁰ (<ftp://ftp.genomics.org.cn/pub/gce>) was then used to estimate the genome size of *M. lutarioriparius* with modified parameters: -c 24, -m 1, -D 8 -H 1. Other parameters were left as default. Genome size of *M. lutarioriparius* was estimated with the formula: genome size = now_kmer/cvg, where now_kmer is the total number of k-mer after filtering the number of k-mer with low coverage and cvg is the estimated average sequencing depth; The repeat content was estimated with the formula: repeat content = $1-b[1/2]-b[1]$, and the heterozygosity was estimated using the formula: heterozygosity = $(a[1/2]/(2-a[1/2]))/k\text{-mer}$.

Genome assembly

Before assembly, raw Oxford Nanopore reads were self-corrected using Canu software⁷¹ (v1.8) with modified parameters as follow: genomeSize = 2G, minReadLength = 2000, minOverlapLength = 500, -nanopore-corrected. After self-correction, a total of 112.78 Gb data (the longest 40× reads) were kept for following assembly. The corrected Nanopore reads were then assembled with SMARTdenovo software (<https://github.com/ruanjue/smartdenovo>) with modified parameters: -k 21, the others were set as default. The initial contig assembly was polished through three iterations using Racon software (<https://github.com/lbcb-sci/racon>) with raw Nanopore reads and default parameters. Contigs were further polished with Pilon⁷² (v1.23) through three iterations using 172.52 Gb filtered paired-end 150 bp Illumina data. The Pilon parameters were modified as following: --fix bases, --changes, --vcf, --diploid. The quality of assembly after each round polish was assessed using BUSCO.

Hi-C scaffolding

To confirm the reliability of the Hi-C scaffolding, two Hi-C scaffolding programs LACHESIS²⁸ (<http://shendurelab.github.io/LACHESIS/>) and 3D *de novo* assembly (3d-dna) pipeline²⁹ (<https://github.com/theaidenlab/3d-dna>) were tested to anchor the contigs in this study, and the better one that generated by LACHESIS was selected as the final assembly.

Before scaffolding, raw Hi-C reads were removed adapters and trimmed low-quality bases using fastp⁶⁵ (version 0.12.6) with default parameters. A total of 4,772 M clean reads were mapped to the contig assembly using BWA⁷³ (version 0.7.17). The percentage of informative Hi-C reads was estimated using HiC-pro⁶⁸ (version 2.10.0) with default parameters. The deduplicated list of alignments of Hi-C reads to the contig assembly was generated using Juicer pipeline⁷⁴ (version 1.5.7). 3d-dna²⁹ (version 180922) was then used to anchor the contig assembly into chromosome-length scaffolds with the linking information contained in Hi-C reads. 3d-dna was tested to run with haploid and diploid mode, respectively. The iterative rounds for mis-correction were tested with 0, 2 and 7 times, respectively. Only these contigs with length longer than 15,000 bps will be taken into the process of cluster, order and orientation. The heatmap for Hi-C interaction was processed using 3d-dna visualize module and reviewed in Juicebox⁷⁵ (version 1.9.0).

LACHESIS was also used to construct chromosome-level assembly. The quality filtered Hi-C

reads were firstly mapped to the Nanopore contigs assembly using the BWA-aln algorithm⁷³ (version 0.7.17). The duplicated reads were removed using MarkDuplicates program in Picard. And then the PreprocessSAMs.pl program was used to further filter the alignment files. Finally the LACHESIS was used to anchor the Nanopore contigs into pseudochromosomes with modified LACHESIS parameters: CLUSTER_N = 19; CLUSTER_MIN_RE_SITES = 22; CLUSTER_MAX_LINK_DENSITY = 2; CLUSTER_NONINFORMATIVE_RATIO = 2; ORDER_MIN_N_RES_IN_TRUN = 10; ORDER_MIN_N_RES_IN_SHREDS = 10; RE_SITE_SQR = GATC in the configure file. The redundancy sequences were collapsed during the Hi-C scaffolding process. The 19 pseudochromosomes of *M. lutarioriparius* constructed by Hi-C technologies were named Chr01 to Chr19 according to the syntenic relationship with sorghum chromosomes.

Genome assembly quality evaluation

The sequence quality of Nanopore contigs were assessed by sequence comparison with finished Bacterial Artificial Chromosome (BAC) sequences, which were sequenced using Sanger sequencing. Sequence comparisons between Nanopore contigs and BAC sequences were performed using MUMmer⁷⁶ (version 3.23). Reads coverage for contigs was calculated by mapping raw Nanopore reads against BAC sequences using minimap2⁷⁷ (version 2.16) with following modified parameters: -x map-ont -a. The integrity of both genome assembly and predicted protein-coding genes were assessed using benchmarking universal single-copy orthologs (BUSCOs) from the embryophyta lineage (BUSCO version: 3.1.0, database version: embryophyta_odb10)⁷⁸. High-quality short reads generated from three Illumina DNA sequencing libraries were mapped to the *M. lutarioriparius* genome assembly using BWA-MEM algorithm⁷³ (version 0.7.17) with default parameters, and the mapping rate was calculated. The quality of genome assembly was further assessed by the mapping RNA sequencing short reads derived from nine different RNA sequencing libraries using HISAT2⁷⁹ (version 2.0.5) with default parameters, and the overall alignment rates for each sample were used to indicate the quality of assembly. Finally, we assessed the *M. lutarioriparius* genome assembly using the LTR Assembly Index (LAI)³⁰, which evaluate the quality of genome assembly through the content of identifiable intact long terminal repeat retrotransposons (LTR-RTs). LTR-FINDER⁸⁰ (version 1.0.7) was used to search LTR-RTs candidates in *M. lutarioriparius* genome assembly with the modified parameters: -D 20000 -d 1000 -L 700 -l 100 -p 20 -C -M 0.9. LTR_retriever⁸¹ was used to accurately identify LTR-RTs and calculate the LTR Assembly Index (LAI) for *M. lutarioriparius* genome assembly using the output of LTR-FINDER.

Repeat analysis

A repeats library of *M. lutarioriparius* genome was *ab initio* constructed using RepeatModeler (open-1.0.11) (<http://www.repeatmasker.org/RepeatModeler>), which automatically execute RECON and RepeatScout for repeats finding. The consensus TE sequences generated by RepeatModeler software were combined with RepBase-20181026 and used as repeats library in RepeatMasker (version open-4.0.7) (<http://www.repeatmasker.org>) for repetitive elements identification in *M. lutarioriparius* genome. A preliminary list of candidate long terminal repeat retrotransposons (LTR-RTs) for *M. lutarioriparius* was generated using LTR_FINDER⁸⁰ (version

1.0.7) with modified parameters: -D 20000 -d 1000 -L 5000 -l 100 -p 20. The identification of high-quality intact LTR-RTs and the calculation of insertion age for intact LTR-RTs were carried out using LTR_retriever⁸¹ with default parameters. *De novo* searches for MITEs used MITE Tracker software⁸² with default parameters. Tandem repeats detection used Tandem Repeats Finder program⁸³ (version 4.09) with modified parameters: 2 7 7 80 10 50 500 -m -f -d. The locations of centromere and telomere were inferred from the outputs generated by Tandem Repeats Finder.

RNA Sequencing data analysis

The RNA sequencing paired-end reads generated from nine samples, including leaf, root, rhizome, spikelet, lateral bud, seedling, upper part of internode, mid part of internode and lower part of internode, were first removed adapters and trimmed low quality bases using fastp⁶⁵ (version 0.12.6) with parameters: -q 25 -l 75. After quality control, 82.34 Gb clean data from nine RNA sequencing libraries were generated for gene prediction and expression analysis. The clean reads were mapped to the *M. lutarioriparius* genome using HISAT2⁷⁹ (version 2.0.5) with default parameters. Following the alignment, SummarizedExperiment object was constructed using R/Bioconductor package tximeta (version 1.4.2) and raw read count matrix for each gene were derived from SummarizedExperiment object using R function counts.

Gene prediction and genome annotation

Before gene prediction, the repetitive sequences of *M. lutarioriparius* genome was masked by RepeatMasker (<http://www.repeatmasker.org/RepeatModeler.html>). The *ab initio* prediction of protein-coding gene was carried out using FGENESH⁸⁴ and augustus⁸⁵. Fgenesh was used for gene prediction with authorized monocots model and modified parameters: -pmrna -scip_prom -scip_term -min_f_exon:16 -min_i_exon:16 -min_t_exon:16 -min_s_exon:90. augustus (version 3.2.3) was running with following parameters: --strand=both --genemodel=partial --gff3=on --species=maize5. For homolog protein-based gene prediction, the proteomes of rice (*Oryza sativa* v7_JGI), maize (*Zea mays* Ensembl-18), sorghum (*Sorghum bicolor* v3.1.1) and *S. spontaneum*⁴⁰ were mapped to the genome assembly using Exonerate⁸⁶ (version 2.2.0) with Protein2Genome model. For RNA sequencing aided gene annotation, the quality filtered RNA reads of nine libraries, including leaf, root, rhizome, spikelet, lateral bud, seedling, upper part of internode, mid part of internode and lower part of internode, were subjected to Trinity⁸⁷ (version 2.1.1) to perform *de novo* transcript sequence reconstruction with default parameters. Redundant transcript sequences were collapsed using CD-HIT^{88,89} (version 4.7) with modified parameters: -c 0.90 -n 9 -d 0 -M 0 -T 0. A total of 356,126 non-redundant transcript sequences, accounting to 297.68 Mb in total, were finally reconstructed. For genome-guided transcript assembly, the RNA sequencing reads were firstly mapped to repeats masked *M. lutarioriparius* genome assembly using HISAT2⁷⁹ (version 2.0.5). StringTie⁹⁰ (v1.3.2d) was then used to reconstruct the transcripts. The PASA pipeline⁹¹ (v2.2.0) was used to construct comprehensive transcripts through integrating transcripts output by Trinity *de novo* assembly and genome-guided assembled transcripts generated by StringTie.

The EVidenceModeler (EVM)⁹¹ (version 1.1.1) was used to integrate the gene prediction results generated by *ab initio* gene prediction software, Exonerate and PASA. According to our understanding to the reliability of different software, we assigned score 5 to the prediction result of

FGENESH, 3 to augustus, 5 to the prediction results of exonerate based on homologous protein, and 10 to the prediction results of PASA based on transcriptome data set in the configuration file of EVM software.

MAKER⁹² (version 2.31.10) was used to add MAKER's quality-control metrics to the EVM integration results for getting more reliable gene prediction results. At least 85% of exons per gene should have evidence support from RNA-seq transcripts or homolog proteins. And these predicted genes with translated protein sequence length less than 50 amino acid would be filtered.

The final filtered protein-coding genes were then subjected to InterProScan⁹³ (version 5.22-61.0), eggNOG-mapper^{36,94} and KOALA (KEGG Orthology And links Annotation)⁹⁵ for genome annotation. InterProScan was run with the modified parameters of “-iprlookup -goterms -pa -f tsv” and GO annotation of predicted genes was extracted from the output generated by InterProScan.

KEGG and GO enrichment analysis of gene sets were performed in R (version 3.5.0) software (<https://www.r-project.org>) against *M. lutarioriparius* genome as reference. Statistical significance was tested using over-representation test⁹⁶ within R package clusterProfiler⁹⁷. Benjamini-Hochberg (BH) method was used to control the false discovery rate (under the threshold of 0.05) in multiple hypothesis testing⁹⁸.

Non-coding RNAs identification

Non-coding RNAs in *M. lutarioriparius* were identified using Rfam (version 14) and Infernal (version 1.1.2) software. The tRNAs identification used tRNAscan-SE 2.0⁹⁹ software. The rRNA identification used barrnap (<https://github.com/tseemann/barrnap>).

Transcript factor identification

The putative protein sequences of *M. lutarioriparius* were subjected to two web tools plantTFDB 5.0¹⁰⁰ and plantTFcat¹⁰¹ for transcript factor and transcriptional regulator prediction. PlantTFDB predicted 4,031 transcription factors in *M. lutarioriparius* genome. PlantTFcat predicted 8,243 transcript factors and transcriptional regulators in *M. lutarioriparius* genome. And 3,977 genes were found to be shared by two prediction results. Finally, we took the prediction results of plantTFDB as the candidate transcript factors of *M. lutarioriparius*.

Genomic comparisons and visualizations

MCScan³⁹ (Python version) ([https://github.com/tanghaibao/jcvi/wiki/MCscan-\(Python-version\)](https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version))) was used for pairwise synteny region search with the LSAT results of *M. lutarioriparius* versus sorghum. Dot-plots for genome pairwise synteny visualization was generated using the command “python -m jcvi.graphics.dotplot” within MCScan (Python version). The synteny pattern of two compared genomes was using the command “python -m jcvi.compara.synteny depth -- histogram”. Microsynteny visualization on gene level was using the command “python -m jcvi.graphics.synteny”.

Estimate of the recent whole genome duplication (WGD) timing in *M. lutarioriparius*

MCscanX³⁹ was used to perform the self-comparison of *M. lutarioriparius* protein-coding

genes. The synonymous substitution values (K_s) of *M. lutarioriparius* versus *M. lutarioriparius* syntenic gene pairs were calculated using KaKs_calculator¹⁰² (version 2.0) with YN model. In total, 94 large collinear blocks with at least 50 homologous gene pairs were identified, among which 28 collinear blocks have similar large values of K_s , roughly ranging from 0.61 to 0.76, whereas the rest 66 collinear blocks have similar but smaller K_s values, roughly ranging from 0.05 to 0.15. The two peaks of the K_s frequency distribution indicate that an ancient gene duplication burst occurred in *M. lutarioriparius* genome prior to the recent WGD. The timing of *M. lutarioriparius* recent WGD was calculated based on the small K_s peak of *M. lutarioriparius* versus *M. lutarioriparius* collinear gene pairs. The synonymous substitutions rate per site per year (r) equaling $6.5e-9$ was applied to the estimation of *M. lutarioriparius* recent WGD.

The genes of *M. lutarioriparius* were classified as singletons, dispersed duplicates, proximal duplicates, tandem duplicates and segmental/WGD duplicates using duplicate_gene_classifier module in MCScanX³⁹ by parsing the all-vs-all BLASTP results. For the definition of proximal duplicates, two genes are in nearby genomic region and separated by less than 10 genes. If a gene has no BLASTP hint, it will be classified as singletons. The priority of assignment for these genes have multiple BLASTP results is WGD/segmental > tandem > proximal > dispersed.

Phylogenetic analysis and divergence time estimation

To investigate the evolutionary history of *M. lutarioriparius*, seven grass subfamilies, *Brachypodium distachyon* v3.1 (Phytozome v12.1), *Oryza sativa* v7_JGI (Phytozome v12.1), *Zea mays* Ensembl-18 (Phytozome v12.1), *Sorghum bicolor* v3.1.1 (Phytozome v12.1), *Setaria italica* v2.2 (Phytozome v12.1), *Saccharum spontaneum*⁴⁰, *Miscanthus lutarioriparius* and one dicot plant *Arabidopsis thaliana* TAIR10 (Phytozome v12.1) were used for gene family construction by using OrthoFinder⁶³ (version 2.2.7) with default parameters. The protein sequences of 122 single-copy orthologs from 8 species were concatenated for the species tree construction. Multiple sequence alignment for these concatenated single-copy orthologs (protein sequences) was carried out using MAFFT¹⁰³ (version: v7.158b) L-INS-i iterative refinement method. The best fit model (JTT+G+F) was selected by using ModelGenerator software¹⁰⁴ (v0.85) based on Bayesian Information Criterion (BIC). Maximum-likelihood trees were constructed using RAXML¹⁰⁵ (8.2.12) and IQ-TREE¹⁰⁶ (version 1.6.3) with 1,000 replicates bootstrap, respectively. BEAST (Bayesian Evolutionary Analysis Sampling Trees) software¹⁰⁷ (v1.8.3) was used to estimate the species divergence times with JTT substitution model and Gamma categories equaling to 4. The calibrated Yule model and strict clock type was set. The length of chain for MCMC was set 10,000,000 and every 1,000 record the parameters. The calibration points used in BESAT were obtained from the TimeTree database (<http://www.timetree.org/>): *O. sativa* versus *S. bicolor* (median time: 48.0 MYA), *S. bicolor* versus *Z. mays* (median time: 11.56 MYA). The visualization of species tree was carried out using FigTree (v1.4.3) (<https://github.com/rambaut/figtree/>).

Gene family expansion and contraction analysis

The gene-family expansion and contraction in *M. lutarioriparius* genome were determined by comparing with those in 7 other species using Computational Analysis of Gene Family Evolution (CAFE)¹⁰⁸ (version 4.2.1). The gene family size for each species used in CAFE was calculated by

OrthoFinder⁶³ (version 2.2.7). The ultra-metric phylogeny tree was reconstructed using r8s¹⁰⁹ (v1.81) recommended by the tutorial of CAFE. The calibration points were obtained from the TimeTree database (<http://www.timetree.org/>). We suspected that the clade of *M. lutarioriparius*, *S. spontaneum*, *S. bicolor* share the same rate of gene family evolution (birth and death) and the rest species had another rate. To better understand the potential function category of each gene family, we used KinFin¹¹⁰ (v1.0.3) to take gene families identified by OrthoFinder, alongside gene functional annotation assigned by InterProScan, to derive rich aggregative annotation for gene families.

Chloroplast genome assembly and annotation

Illumina short reads generated from whole genome sequencing were used for chloroplast genome assembly using MITObim software¹¹¹ (version 1.9) with default parameters. We used two different bait sequences, including chloroplast genome of *Miscanthus junceus* (NCBI Reference Sequence: LN869216.1) and *Saccharum spontaneum* (NCBI Reference Sequence: LN896360.1) to verify whether different bait chloroplast genome sequences could affect the final assembly results. The GeSeq¹¹² (<https://chlorobox.mpimp-golm.mpg.de/geseq.html>) was used to annotate protein-coding genes, rRNA genes, and tRNA genes in the *M. lutarioriparius* chloroplast genome.

Phylogeny analysis of genus *Miscanthus* based on chloroplast genome sequence

We downloaded the 31 previously released *Miscanthus* chloroplast genomes from NCBI database (Supplementary table 23), including *M. floridulus* (4), *M. junceus* (2), *M. lutarioriparius* (1), *M. sacchariflorus* (6), *M. sinensis* (14), *M. transmorrisonensis* (2), *M. × giganteus* (3) (The number in the parentheses indicates the number of sample). Four chloroplast genomes were used as outgroup, including a *Saccharum spontaneum* (NCBI Reference Sequence: LN896360.1), a maize (NCBI Reference Sequence: NC001666.2), a rice (NCBI Reference Sequence: NC_031333.1) and a *B. distachyon* (NCBI Reference Sequence: NC_011032.1). All chloroplast genome sequences were used to perform multiple sequences alignment using MAFFT¹⁰³ (version: v7.158b) with --auto option. Conserved blocks from the alignment file generated above were evaluated and selected using GBLOCKS¹¹³ (version 0.91b). Finally, 140,678 positions (69 %) in 1,480 blocks were used in the subsequent phylogenetic analysis. IQ-TREE¹⁰⁶ (version 1.6.3) was used to reconstruct phylogenetic tree with maximum likelihood method with 1,000 replicates of ultrafast bootstrap. The best-fit model for molecular evolution was determined according to the Bayesian Information Criterion (BIC) calculated by ModelFinder¹¹⁴. Bayesian inference (BI) of phylogeny was carried out using MrBayes¹¹⁵ (version 3.1.2) with the modified parameters. The number of generations was set to 10,000,000, and the sample frequency of Markov chain was set to 100 generations; Nst was set to 6 and rates was set to Gamma during the execution of MrBayes; *Saccharum spontaneum* (NCBI Reference Sequence: LN896360.1) was set as outgroup; 25% of the samples were discard as burn-in. For neighbor-joining phylogenetic tree reconstruction, MEGAX¹¹⁶ was used with 1,000 bootstrap trails. Time-calibrated phylogeny analysis based on the chloroplast genome sequences was carried out using BEAST (Bayesian Evolutionary Analysis Sampling Trees) software¹⁰⁷ (v1.8.3). Sites model was set as HKY, and site heterogeneity model was set as Gamma + Invariant sites. Number of gamma categories was set as 4. Length of chain was set as 10,000,000. Log

parameters every 1,000. Only one calibrations point from TimeTree (<http://www.timetree.org/>) was used (*Saccharum spontaneum* versus *Miscanthus lutarioriparius* median time: 6.98 MYA, Range: 5.53-8.43 MYA). FigTree (v1.4.3) (<https://github.com/rambaut/figtree/>) was used to draw the dendrograms.

Identification of plant disease resistance genes (R genes)

We used DRAGO 2 tool of Plant Resistance Genes database (PRGdb 3.0, <http://prgdb.org/prgdb/genes>)¹¹⁷ to predict and annotate candidate Pathogen Receptor Genes (PRGs) with high sensitivity and accuracy. DRAGO 2 tool is an integrated pipeline for R domains identification, which can also identify both coiled-coil (CC) structure in the sequences and TM signal using Coils 2.2 program¹¹⁸ and TMHMM 2c program¹¹⁹ respectively. The R genes of *M. lutarioriparius* genome were further classified into CNL (CC-NB-LRR), TNL (TIR-NB-LRR), RLP (ser/thr-LRR), RLK (Kin-LRR) and others based on the presence of specific domains.

We also employed HMMER software¹²⁰ (version 3.1 b1-linux-intel - x86_64) to identify NBS-LRR genes in *M. lutarioriparius* using Hidden Markov Model (HMM) profile for Pfam¹²¹ NBS domain (NB-ARC, PF00931, <http://pfam.xfam.org/>). A high-quality protein set (E-value < 1×10^{-4}) was kept as candidate for further filtration. In addition, 231 reference sequences of NBS-LRR genes were download from NCBI and used for BLASTP¹²² (version 2.2.28+) search. The E-value for BLAST was set as 0.01.

Only these genes identified by both DRAGO 2 tool and HMMER software would be taken as candidate NBS-LRR genes. The proteins of candidate NBS-LRR genes were then subjected to Batch Web CD-Search (<https://www.ncbi.nlm.nih.gov/Structure/>) for function domain annotation. The final candidate NBS-LRR genes were filtered by manual check according to the results of Batch Web CD-Search and BLASTP. The NBS-associated conserved domains TIR (PF01582) and RPW8 (PF05659) were identified using HMMER software with default parameters. While RPW8 domain didn't present in our *M. lutarioriparius* genome.

Identification and phylogeny analysis of cell wall biosynthesis related gene families in *M. lutarioriparius*

Known cellulose synthases (CesA) and cellulose synthase-like (Csl) coding sequences of rice, maize and *Arabidopsis* were downloaded from the Cell Wall genomics website (<https://cellwall.genomics.purdue.edu/families/>). To identified the candidate CesA and Csl genes in *M. lutarioriparius* genome, we performed BLAST¹²² (version 2.2.28+) search against the coding sequences of *M. lutarioriparius* with these known cellulose synthases and cellulose synthase-like protein sequences. The E-value threshold for BLAST search was set to $1e^{-5}$. The BLAST hints for CesA gene identification were manually filtered based on that the cumulative length of same target gene should not less than 1,300 bp.

HMM profiles for the CesA and Csl family (PF03552, Cellulose_synt and PF00535, Glycos_transf_2 or GT2) obtained from the Pfam database¹²¹ (<http://pfam.xfam.org/>) were used to search against *M. lutarioriparius* protein sequences using HMMER¹²⁰ (3.1b1-linux-intel-x86_64) with cut_tc mode. The E-value cutoff was set to $1e^{-3}$. We kept the results detected by both HMMER and BLAST search, which means every candidate protein sequence should be similar to the BLAST

query CesA and Csl and own either of the two Pfam domains.

We took the collinearity with sorghum CesA genes into the identification of CesA in *Miscanthus*. Twelve CesA related genes of sorghum were functional annotated by Phytozome v12.1 (<https://phytozome.jgi.doe.gov/>). 28 candidate CesA genes were identified in *M. lutarioriparius* based on the genome collinearity between *M. lutarioriparius* and sorghum. One of these 28 candidate genes was removed due to the coding sequence length (MI02G069810: 429 bp). The rest 27 genes were well supported by both HMMER and BLAST search above.

For phylogenetic analysis of CesA and Csl genes, the candidate CesA and Csl protein sequences were firstly performed multiple sequence alignment using MAFFT¹⁰³ (version: v7.158b) with L-INS-i method. The maximum likelihood tree was reconstructed using IQ-TREE¹⁰⁶ (version 1.6.3) with the best-fit model automatically selected by ModelFinder. 1,000 replicates of ultrafast bootstrap were carried out. The best-fit model chosen according to BIC was JTT+F+R6. Gene structure information was parsed from the genome annotation file and displayed by using Gene Structure Display Server (GSDS 2.0) (<http://gsds.cbi.pku.edu.cn>).

Carbohydrate-active enzymes (CAZyme) are important type of proteins responsible for the synthesis, degradation and modification of carbohydrates. The CAZy database (CAZyDB)^{56,57} (<http://www.cazy.org>) classified CAZyme families into five functional types: glycosyltransferases (GTs), glycoside hydrolases (GHs), carbohydrate esterases (CEs), polysaccharide lyases (PLs) and the non-catalytic carbohydrate-binding modules (CBMs). Here, we used dbCAN2 meta server (<http://bcb.unl.edu/dbCAN2/>) to annotate carbohydrate-active enzyme in *M. lutarioriparius*. dbCAN2 meta server integrates three state-of-art tools HMMER, DIAMOND and Hotpep for automated CAZyme annotation (DIAMOND: E-Value < 1e-102, hits per query (-k) = 1; HMMER: E-Value < 1e-15, coverage > 0.35; Hotpep: Frequency > 2.6, Hits > 6). CAZymes are defined as those predicted by at least 2 tools mentioned above.

Reference

1. Chen, S. L. & Renvoize, S. A. A new species and a new combination of *Miscanthus* (Poaceae) from China. *Kew Bull.* **60**, 605–607 (2005).
2. Tang, Y. M. et al. Molecular cytogenetic characterization and phylogenetic analysis of four *Miscanthus* species (Poaceae). *Comp. Cytogenet.* **13**, 211–230 (2019).
3. Heaton, E. A. et al. *Miscanthus*: a promising biomass crop. *Adv. Bot. Res.* **56**, 75–137 (2010).
4. Malinská et al. Physiological response of *Miscanthus* × *giganteus* to plant growth regulators in nutritionally poor soil. *Plants* **9**, 194 (2020).
5. Głowacka, K. et al. Can chilling tolerance of C₄ photosynthesis in *Miscanthus* be transferred to sugarcane? *GCB Bioenergy* **8**, 407–418 (2016).
6. Wang, D., Naidu, S. L., Portis, A. R., Moose, S. P. & Long, S. P. Can the cold tolerance of C₄ photosynthesis in *Miscanthus* × *giganteus* relative to *Zea mays* be explained by differences in activities and thermal properties of Rubisco? *J. Exp. Bot.* **59**, 1779–1787 (2007).
7. Zadel, U. et al. Changes induced by heavy metals in the plant-associated microbiome of *Miscanthus* × *giganteus*. *Sci. Total Environ.* **711**, 134433 (2020).
8. Pogrzeba, M., Rusinowski, S. & Krzyżak, J. Macroelements and heavy metals content in energy crops cultivated on contaminated soil under different fertilization—case studies on autumn harvest. *Environ. Sci. Pollut. Res.* **25**, 12096–12106 (2018).

-
- 855 9. Barbosa, B. et al. Phytoremediation of heavy metal-contaminated soils using the perennial
856 energy crops *Miscanthus* spp. and *Arundo donax* L. *Bioenergy Res.* **8**, 1500–1511 (2015).
 - 857 10. Ezaki, B., Nagao, E., Yamamoto, Y., Nakashima, S. & Enomoto, T. Wild plants, *Andropogon*
858 *virginicus* L. and *Miscanthus sinensis* Anders, are tolerant to multiple stresses including
859 aluminum, heavy metals and oxidative stresses. *Plant Cell Rep.* **27**, 951–961 (2008).
 - 860 11. Pidlisnyuk, V., Stefanovska, T., Lewis, E. E., Erickson, L. E. & Davis, L. C. *Miscanthus* as a
861 productive biofuel crop for phytoremediation. *Crit. Rev. Plant Sci.* **33**, 1–19 (2014).
 - 862 12. Bang, J. et al. Phytoremediation of heavy metals in contaminated water and soil using
863 *Miscanthus* sp. Goedae-Uksae 1. *Int. J. Phytoremediation* **17**, 515–520 (2015).
 - 864 13. Arduini, I., Masoni, A., Mariotti, M. & Ercoli, L. Low cadmium application increase
865 miscanthus growth and cadmium translocation. *Environ. Exp. Bot.* **52**, 89–100 (2004).
 - 866 14. Ho, C.P., Hseu, Z.Y., Chen, N.C. & Tsai, C.C. Evaluating heavy metal concentration of plants
867 on a serpentine site for phytoremediation applications. *Environ. Earth Sci.* **70**, 191–199 (2013).
 - 868 15. Guo, H. et al. Malate secretion from the root system is an important reason for higher
869 resistance of *Miscanthus sacchariflorus* to cadmium. *Physiol. Plant.* **159**, 340–353 (2017).
 - 870 16. Chen, S. & Renvoize, S. A. *Miscanthus*. *Flora of China* **22**, 581–583 (2006).
 - 871 17. Liu, C. et al. Biomass properties from different *Miscanthus* species. *Food Energy Secur.* **2**,
872 12–19 (2013).
 - 873 18. Li, X. et al. Nuclear DNA content variation of three *Miscanthus* species in China. *Genes and*
874 *Genomics* **35**, 13–20 (2013).
 - 875 19. Wang, Q. et al. Transcriptomic evaluation of *Miscanthus* photosynthetic traits to salinity
876 stress. *Biomass and Bioenergy* **125**, 123–130 (2019).
 - 877 20. Zhu, C. et al. Haplotypes phased from population transcriptomes detecting selection in the
878 initial adaptation of *Miscanthus lutarioriparius* to stressful environments. *The Plant Genome* **10**,
879 plantgenome2016.11.0119 (2017).
 - 880 21. Liu, W. & Sang, T. Potential productivity of the *Miscanthus* energy crop in the Loess Plateau
881 of China under climate change. *Environ. Res. Lett.* **8**, 044003 (2013).
 - 882 22. Takayama, S. & Isogai, A. Self-incompatibility in plants. *Annu. Rev. Plant Biol.* **56**, 467–489
883 (2005).
 - 884 23. Rayburn, A. L., Crawford, J., Rayburn, C. M. & Juvik, J. A. Genome size of three
885 *Miscanthus* species. *Plant Mol. Biol. Report.* **27**, 184–188 (2009).
 - 886 24. Chae, W. B. et al. Plant morphology, genome size, and SSR markers differentiate five distinct
887 taxonomic groups among accessions in the genus *Miscanthus*. *GCB Bioenergy* **6**, 646–660 (2014).
 - 888 25. Nishiwaki, A. et al. Discovery of natural *Miscanthus* (Poaceae) triploid plants in sympatric
889 populations of *Miscanthus sacchariflorus* and *Miscanthus sinensis* in southern Japan. *Am. J. Bot.*
890 **98**, 154–159 (2011).
 - 891 26. Dwiyanti, M. S. et al. Genetic analysis of putative triploid *Miscanthus* hybrids and tetraploid
892 *M. sacchariflorus* collected from sympatric populations of Kushima, Japan. *BioEnergy Res.* **6**,
893 486–493 (2013).
 - 894 27. Wei, S., Yang, Y. & Yin, T. The chromosome-scale assembly of the willow genome provides
895 insight into Salicaceae genome evolution. *Hortic. Res.* **7**, 45 (2020).
 - 896 28. Burton, J. N. et al. Chromosome-scale scaffolding of *de novo* genome assemblies based on
897 chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
 - 898 29. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields

899 chromosome-length scaffolds. *Science* **356**, 92–95 (2017).

900 30. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly
901 Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).

902 31. Richards, E. J. & Ausubel, F. M. Isolation of a higher eukaryotic telomere from *Arabidopsis*
903 *thaliana*. *Cell* **53**, 127–136 (1988).

904 32. Zoubak, S., Clay, O. & Bernardi, G. The gene distribution of the human genome. *Gene* **174**,
905 95–102 (1996).

906 33. Gardiner, K. Base composition and gene distribution: critical patterns in mammalian genome
907 organization. *Trends Genet.* **12**, 519–524 (1996).

908 34. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–
909 921 (2001).

910 35. Quevillon, E. et al. InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–
911 W120 (2005).

912 36. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically
913 annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**,
914 D309–D314 (2019).

915 37. Swaminathan, K. et al. A framework genetic map for *Miscanthus sinensis* from RNAseq-
916 based markers shows recent tetraploidy. *BMC Genomics* **13**, 142 (2012).

917 38. Rizzon, C., Ponger, L. & Gaut, B. S. Striking similarities in the genomic distribution of
918 tandemly arrayed genes in *Arabidopsis* and rice. *PLoS Comput. Biol.* **2**, e115 (2006).

919 39. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny
920 and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).

921 40. Zhang, J. et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum*
922 *spontaneum* L. *Nat. Genet.* **50**, 1565–1573 (2018).

923 41. van der Biezen, E. A. & Jones, J. D. G. The NB-ARC domain: a novel signalling motif
924 shared by plant resistance gene products and regulators of cell death in animals. *Curr. Biol.* **8**,
925 R226–R228 (1998).

926 42. McHale, L., Tan, X., Koehl, P. & Micheltore, R. W. Plant NBS-LRR proteins: adaptable
927 guards. *Genome Biology* **7**, 212 (2006).

928 43. Meyers, B. C., Kozik, A., Griego, A., Kuang, H. & Micheltore, R. W. Genome-wide
929 analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* **15**, 809–834 (2003).

930 44. Leister, D. Tandem and segmental gene duplication and recombination in the evolution of
931 plant disease resistance genes. *Trends Genet.* **20**, 116–122 (2004).

932 45. Millet, Y. A. et al. Innate immune responses activated in *Arabidopsis* roots by microbe-
933 associated molecular patterns. *Plant Cell* **22**, 973–990 (2010).

934 46. Schuler, M. A. The role of cytochrome P450 monooxygenases in plant-insect interactions.
935 *Plant Physiol.* **112**, 1411–1419 (1996).

936 47. Minibayeva, F., Beckett, R. P. & Kranner, I. Roles of apoplastic peroxidases in plant response
937 to wounding. *Phytochemistry* **112**, 122–129 (2015).

938 48. Sofo, A., Scopa, A., Nuzzaci, M. & Vitti, A. Ascorbate peroxidase and catalase activities and
939 their genetic regulation in plants subjected to drought and salinity stresses. *Int. J. Mol. Sci.* **16**,
940 13561–13578 (2015).

941 49. Almagro, L. et al. Class III peroxidases in plant defence reactions. *J. Exp. Bot.* **60**, 377–390
942 (2009).

50. Moschetti, I. et al. Constitutive expression of the xylanase inhibitor TAXI-III delays fusarium head blight symptoms in durum wheat transgenic plants. *Mol. Plant-Microbe Interact.* **26**, 1464–1472 (2013).

51. Dangl, J. L. & Jones, J. D. G. Plant pathogens and integrated defence responses to infection. *Nature* **411**, 826–833 (2001).

52. Yang, S., Zhang, X., Yue, J. X., Tian, D. & Chen, J. Q. Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol. Genet. Genomics* **280**, 187–198 (2008).

53. Mace, E. et al. The plasticity of NBS resistance genes in sorghum is driven by multiple evolutionary processes. *BMC Plant Biol.* **14**, 253 (2014).

54. Paterson, A. H. et al. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).

55. Cheng, X. et al. A genomic analysis of disease-resistance genes encoding nucleotide binding sites in *Sorghum bicolor*. *Genet. Mol. Biol.* **33**, 292–297 (2010).

56. Yin, Y. et al. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40**, W445–W451 (2012).

57. Zhang, H. et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101 (2018).

58. Yonekura-Sakakibara, K. & Hanada, K. An evolutionary view of functional diversity in family 1 glycosyltransferases. *Plant J.* **66**, 182–193 (2011).

59. Cheng, H. et al. Metal (Pb, Zn and Cu) uptake and tolerance by mangroves in relation to root anatomy and lignification/suberization. *Tree Physiol.* **34**, 646–656 (2014).

60. Chiniquy, D. et al. XAX1 from glycosyltransferase family 61 mediates xylosyltransfer to rice xylan. *Proc. Natl. Acad. Sci.* **109**, 17117–17122 (2012).

61. Hodkinson, T. R. et al. The use of dna sequencing (ITS and trnL-F), AFLP, and fluorescent in situ hybridization to study allopolyploid *Miscanthus* (Poaceae). *Am. J. Bot.* **89**, 279–286 (2002).

62. Ma, X.F. et al. High resolution genetic mapping by genome sequencing reveals genome duplication and tetraploid genetic structure of the diploid *Miscanthus sinensis*. *PLoS One* **7**, e33821 (2012).

63. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).

64. Doyle, J. & Doyle, J. A rapid isolation procedure for small amounts of leaf tissue. *Phytochem. Bull.* **28**, 330 (1987).

65. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

66. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

67. van Berkum, N. L. et al. Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* **8**, 1937 (2010).

68. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).

69. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

70. Liu, B. et al. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. *arXiv*: **13092012** (2013).

987 71. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting
988 and repeat separation. *Genome Res.* **27**, 722–736 (2017).

989 72. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection
990 and genome assembly improvement. *PLoS One* **9**, e112963 (2014).

991 73. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows-wheeler transform.
992 *Bioinformatics* **25**, 1754–1760 (2009).

993 74. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C
994 experiments. *Cell Syst.* **3**, 95–98 (2016).

995 75. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with
996 unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).

997 76. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**,
998 R12 (2004).

999 77. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–
1000 3100 (2018).

1001 78. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M.
1002 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.
1003 *Bioinformatics* **31**, 3210–3212 (2015).

1004 79. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory
1005 requirements. *Nat. Methods* **12**, 357–360 (2015).

1006 80. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR
1007 retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).

1008 81. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification
1009 of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).

1010 82. Crescente, J. M., Zavallo, D., Helguera, M. & Vanzetti, L. S. MITE Tracker: an accurate
1011 approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC*
1012 *Bioinformatics* **19**, 348 (2018).

1013 83. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*
1014 **27**, 573–580 (1999).

1015 84. Solovyev, V., Kosarev, P., Seledsov, I. & Vorobyev, D. Automatic annotation of eukaryotic
1016 genes, pseudogenes and promoters. *Genome Biol.* **7 Suppl 1**, S10.1-12 (2006).

1017 85. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes
1018 with a generalized hidden Markov model that uses hints from external sources. *BMC*
1019 *Bioinformatics* **7**, 62 (2006).

1020 86. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence
1021 comparison. *BMC Bioinformatics* **6**, 1–11 (2005).

1022 87. Haas, B. J. et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity
1023 platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).

1024 88. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and
1025 comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).

1026 89. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein
1027 or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

1028 90. Kovaka, S. et al. Transcriptome assembly from long-read RNA-seq alignments with
1029 StringTie2. *Genome Biol.* **20**, 278 (2019).

1030 91. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler

and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).

92. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinforma.* **48**, (2014).

93. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

94. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).

95. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726–731 (2016).

96. Boyle, E. I. et al. GO::TermFinder--open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710–3715 (2004).

97. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omi. A J. Integr. Biol.* **16**, 284–287 (2012).

98. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

99. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *bioRxiv* (2019). doi:10.1101/614032

100. Jin, J. et al. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* **45**, D1040–D1045 (2017).

101. Dai, X., Sinharoy, S., Udvardi, M. & Zhao, P. PlantTFcat: an online plant transcription factor and transcriptional regulator categorization and analysis tool. *BMC Bioinformatics* **14**, 321 (2013).

102. Zhang, Z. et al. KaKs_Calculator: calculating *Ka* and *Ks* through model selection and model averaging. *Genomics. Proteomics Bioinformatics* **4**, 259–263 (2006).

103. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

104. Keane, T. M. et al. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* **6**, 29 (2006).

105. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

106. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

107. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).

108. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).

109. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301–302 (2003).

110. Laetsch, D. R. & Blaxter, M. L. KinFin: software for taxon-aware analysis of clustered protein sequences. *G3 Genes, Genomes, Genet.* **7**, 3349–3357 (2017).

111. Hahn, C., Bachmann, L. & Chevreux, B. Reconstructing mitochondrial genomes directly

from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res.* **41**, e129–e129 (2013).

112. Tillich, M. et al. GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11 (2017).

113. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).

114. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermini, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

115. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).

116. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

117. Osuna-Cruz, C. M. et al. PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Res.* **46**, D1197–D1201 (2018).

118. Lupas, A. [30] Prediction and analysis of coiled-coil structures. *Methods Enzymol.* **266**, 513–524 (1996).

119. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).

120. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).

121. Finn, R. D. et al. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).

122. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

Acknowledgements

We thank Qing Gao and Jinxing Chen from CAS Center of Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology for greenhouse management. We thank Ying Lu from college of fisheries and life science, Shanghai Ocean University for valuable advice on genome annotation. We thank Zhen Li, Danfeng Lyv, Minhuan Zeng, Lin Weng, Li Liu from CAS Center of Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology for valuable suggestions. This work was supported by grant from the Chinese Academy of Sciences (XDB27010301 and KSCX2-YW-G-034 to B.H.) and the National Natural Science Foundation of China (31788103).

Author contributions

B.H. and T.S. conceived the project and its components. J.M., Q.F. and B.H. designed the studies and experiments. Q.F., J.Y. and J.M. contributed to the collection and cultivation of *Miscanthus* samples. J.Y. provided the genome size estimation by flow cytometry. C.Z., D.F., Yiqi Lu, Q.T., W.L., Q.W. and Q.F. performed DNA and RNA preparation for genome and transcriptome

1115 sequencing. J.M., H.L. and Q.Z. performed *de novo* genome assembly. L.Z. performed BAC
1116 sequence assembly and sequencing data preprocessing. J.M. and Y.L. performed genome data
1117 analysis and transcriptome analyses. Y.Z. and T.H. provided IT support. L.L., X.H. supervised the
1118 project. J.M., Q.F. and B.H. analyzed the whole data and wrote the manuscript.

Figures

Figure 1

Overview of the *M. lutarioriparius* genome assembly.

Figure 2

Whole genome duplication events within *M. lutarioriparius* genome

Figure 3

Comparative genomics of gene family of *M. lutarioriparius*

Figure 4

Time-calibrated phylogeny analysis based on the chloroplast genomes

Supplementary Files

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

- [Miscanthus20200921SupFigurecompact.pdf](#)
- [MSMiscanthus20200921SupTableMJS.docx](#)
- [MSMiscanthus20200921SupplementaryNotecomact2.docx](#)