

Supplementary Figures

Chromosome-scale assembly and analysis of biomass crop

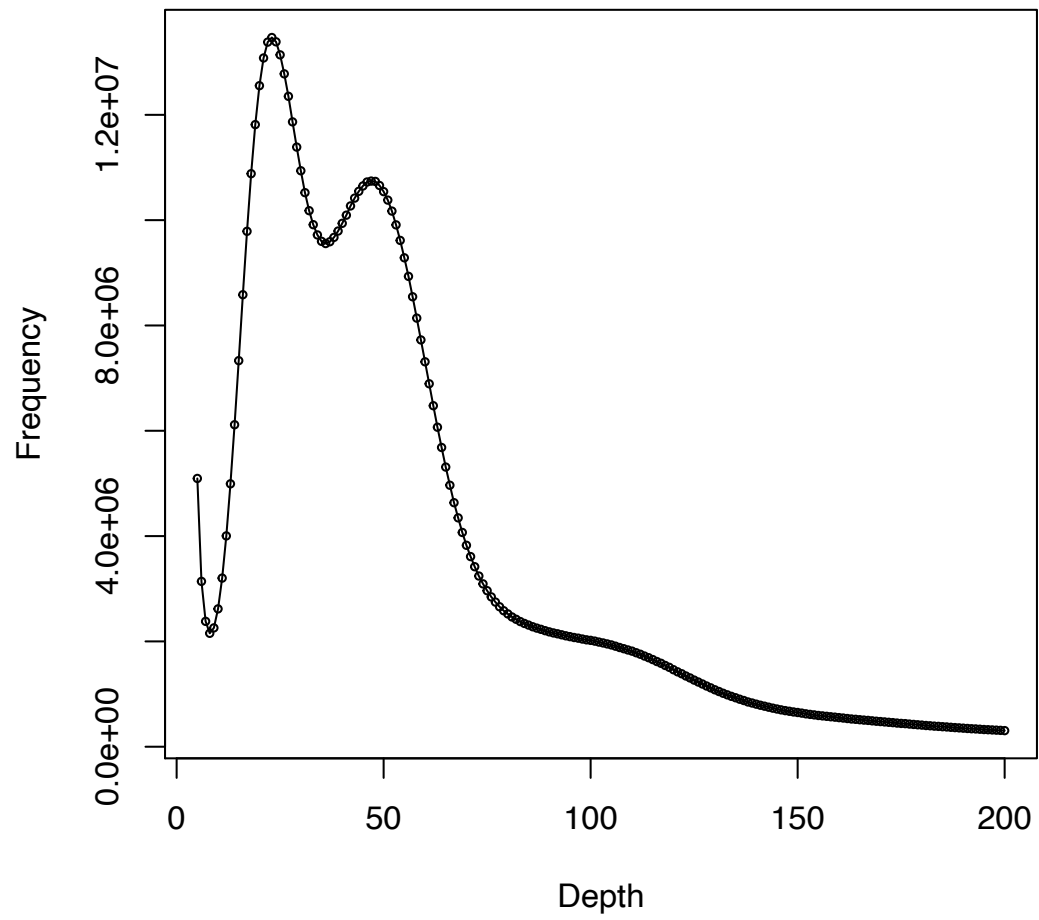
Miscanthus lutarioriparius genome

Supplementary Fig. 1	The 17-mer frequency distribution using NGS short reads. . .	1
Supplementary Fig. 2	The length distribution of filtered Oxford Nanopore reads. . .	2
Supplementary Fig. 3	Sequence comparison between Nanopore contigs and BAC sequences of <i>M. lutarioriparius</i>	3
Supplementary Fig. 4	Interaction frequency distribution of Hi-C links among chromosomes.	4
Supplementary Fig. 5	Comparison of sequences identity of chromosome anchoring results generated by LACHESIS and 3d-dna software.	5
Supplementary Fig. 6	Comparison of whole genome sequence identity between <i>M. lutarioriparius</i> LACHESIS genome assembly and sorghum genome.	6
Supplementary Fig. 7	Boxplot of LTR Assembly Index (LAI) for 19 chromosomes.	7
Supplementary Fig. 8	GC content of each chromosome of <i>M. lutarioriparius</i>	7
Supplementary Fig. 9	Potential centromere and telomere position of <i>M. lutarioriparius</i> genome.	8
Supplementary Fig. 10	The length distribution of protein-coding sequence (CDS), exon and intron of <i>M. lutarioriparius</i> , <i>S. spontaneum</i> and <i>S. bicolor</i>	9
Supplementary Fig. 11	Comparison of GC content distribution (a) and GC3s content distribution (b) of whole genome coding sequences (CDS) for 8 species.	10
Supplementary Fig. 12	The correlation analysis of GC3s, GC content and effective number of codons (Nc).	11
Supplementary Fig. 13	Correlation analysis of GC content and gene density.	12
Supplementary Fig. 14	The frequency distribution of intact LTR-RTs insertion time.	12
Supplementary Fig. 15	The density distribution of repeat elements across 19 chromosomes	13
Supplementary Fig. 16	Distribution of divergence rates of different transposable element (TE) types in <i>M. lutarioriparius</i> genome.	14

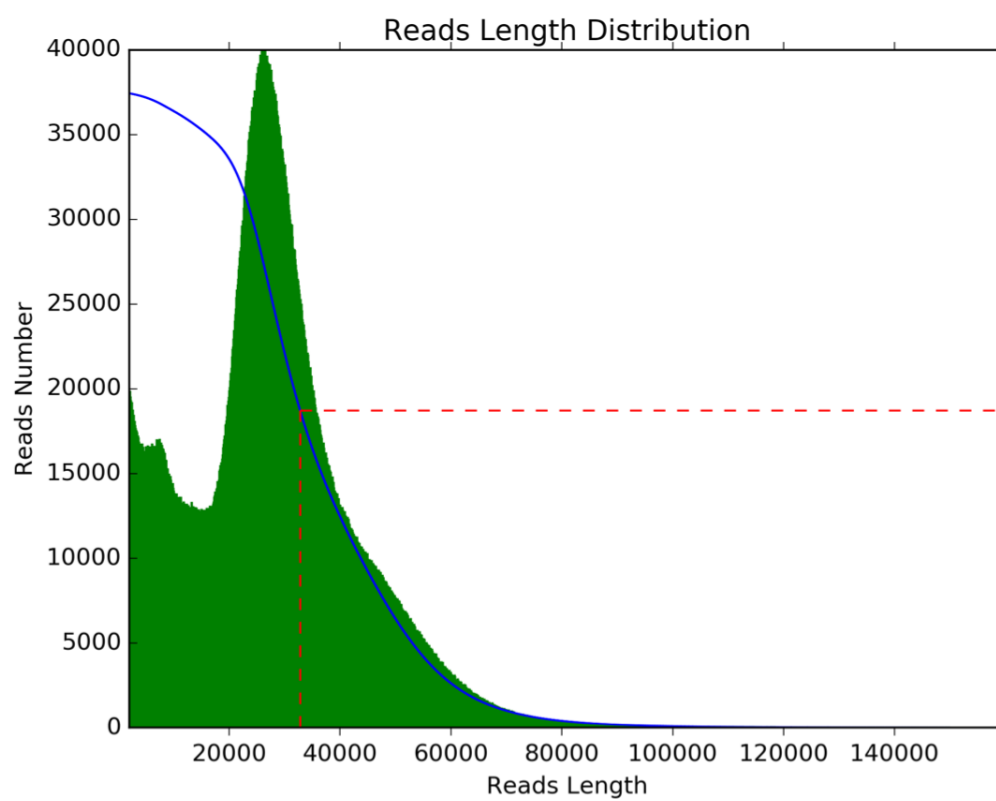
Supplementary Fig. 17 The density distribution of MITEs.....	15
Supplementary Fig. 18 Different origins of genes in <i>M. lutarioriparius</i> genome...	16
Supplementary Fig. 19 The distribution of GC3s, effective number of codons (ENC/Nc) and GC content for genes resulted from different duplication mode.	17
Supplementary Fig. 20 Gene Ontology (GO) enrichment analysis of <i>M.</i> <i>lutarioriparius</i> WGD/segmental duplication.....	18
Supplementary Fig. 21 Enrichment Map for enrichment result of over-representation test for gene duplicates from tandem duplication.	19
Supplementary Fig. 22 Gene Ontology (GO) enrichment analysis of <i>M.</i> <i>lutarioriparius</i> genes resulted from tandem duplication (molecular function category).	20
Supplementary Fig. 23 Gene Ontology (GO) enrichment analysis of <i>M.</i> <i>lutarioriparius</i> gene duplicates from proximal duplication.....	21
Supplementary Fig. 24 Gene Ontology (GO) enrichment analysis of <i>M.</i> <i>lutarioriparius</i> gene duplicates of dispersed duplication origin.	22
Supplementary Fig. 25 Dotplot of GO enrichment for 9,509 expanded gene families of <i>M. lutarioriparius</i>	23
Supplementary Fig. 26 Phylogenetic tree reconstructed based on the amino acid sequences in the NBS domain of NBS-encoding genes from <i>M. lutarioriparius</i>	24
Supplementary Fig. 27 CAZyme domain classes frequency distribution in twelve plant species	25
Supplementary Fig. 28 Frequency distribution of gene members of GT family ...	26
Supplementary Fig. 29 CesA genes of <i>M. lutarioriparius</i>	26
Supplementary Fig. 30 Csl gene family of <i>M. lutarioriparius</i>	27
Supplementary Fig. 31 The marcosynteny of CslF family among <i>M. lutarioriparius</i> , rice and sorghum.	28
Supplementary Fig. 32 Genes involved in lignin biosynthesis in <i>M. lutarioriparius</i>	29
Supplementary Fig. 33 Visualization of annotation of <i>M. lutarioriparius</i> chloroplast genome assembled in this study	30
Supplementary Fig. 34 Phylogenetic tree of genus <i>Miscanthus</i> reconstructed using Maximum Likelihood (ML) approach based on whole chloroplast genome sequence.....	31

Supplementary Fig. 35 Phylogenetic tree of genus *Miscanthus* reconstructed using Neighbor-Joining (NJ) method based on whole chloroplast genome sequence. 32

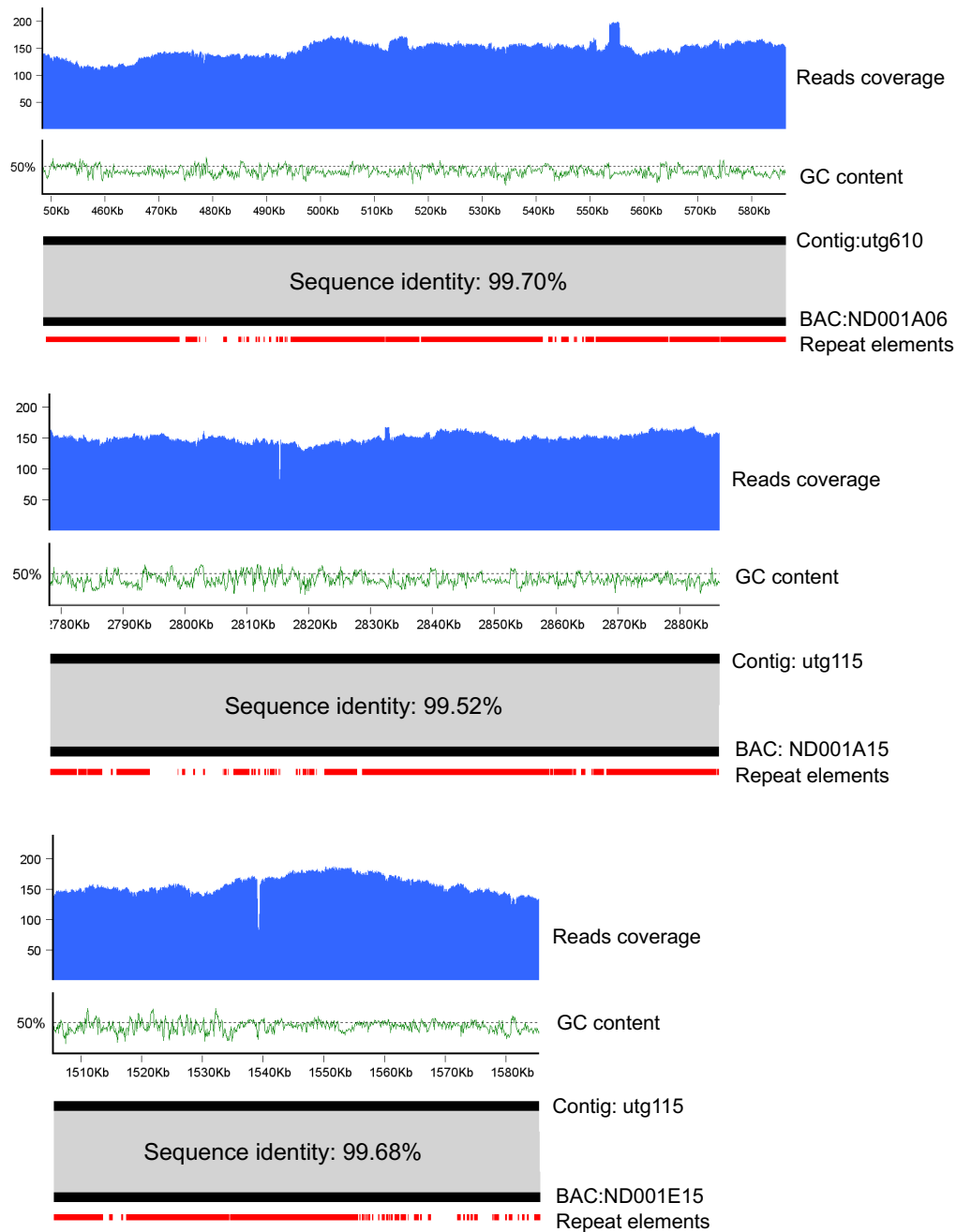
Kmer distribution



Supplementary Fig. 1 The 17-mer frequency distribution using NGS short reads.

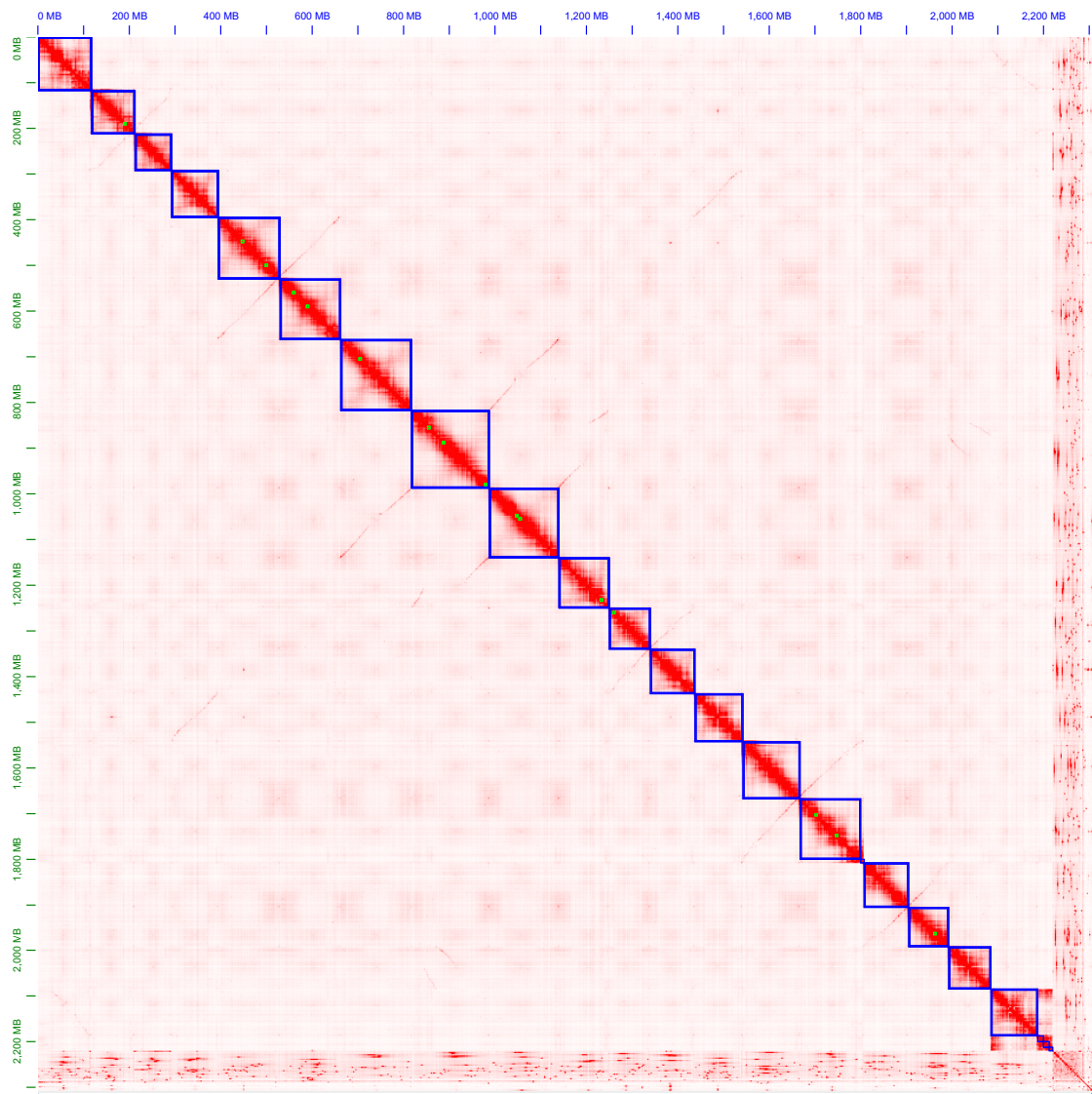


Supplementary Fig. 2 The length distribution of filtered Oxford Nanopore reads.



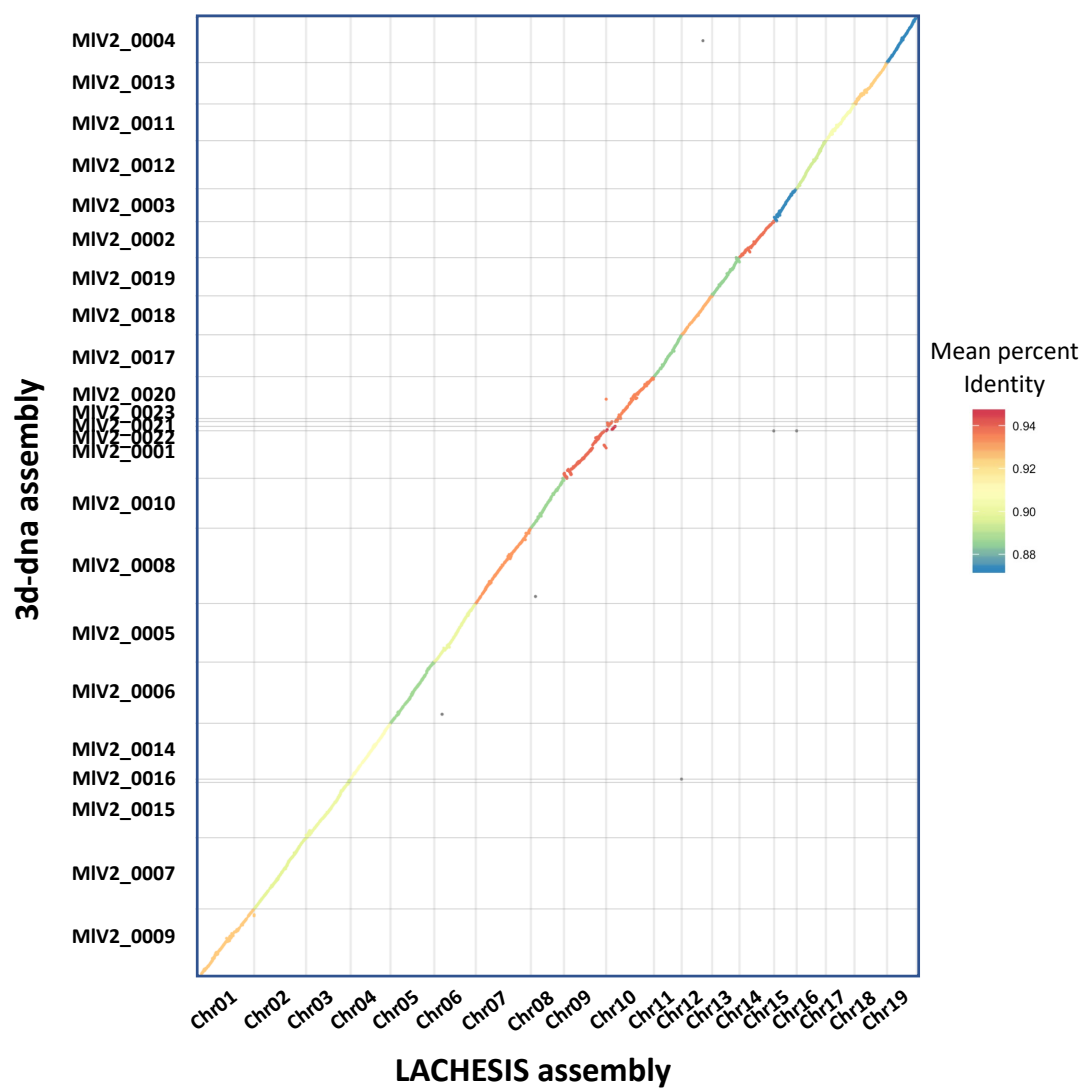
Supplementary Fig. 3 Sequence comparison between Nanopore contigs and BAC sequences of *M. lutarioriparius*

Reads coverage was calculated by mapping the filtered Oxford Nanopore reads against the polished contigs of *M. lutarioriparius* and using a 100 bp window size. GC content was calculated using a 100 bp window size. The repeat elements of BAC sequences were identified using RepeatMasker with *Miscanthus* specific repeats database generated in present study. The sequence comparisons were performed by MUMer 3.2 software. The sequence identity was calculated using the results generated by MUMer 3.2 software.



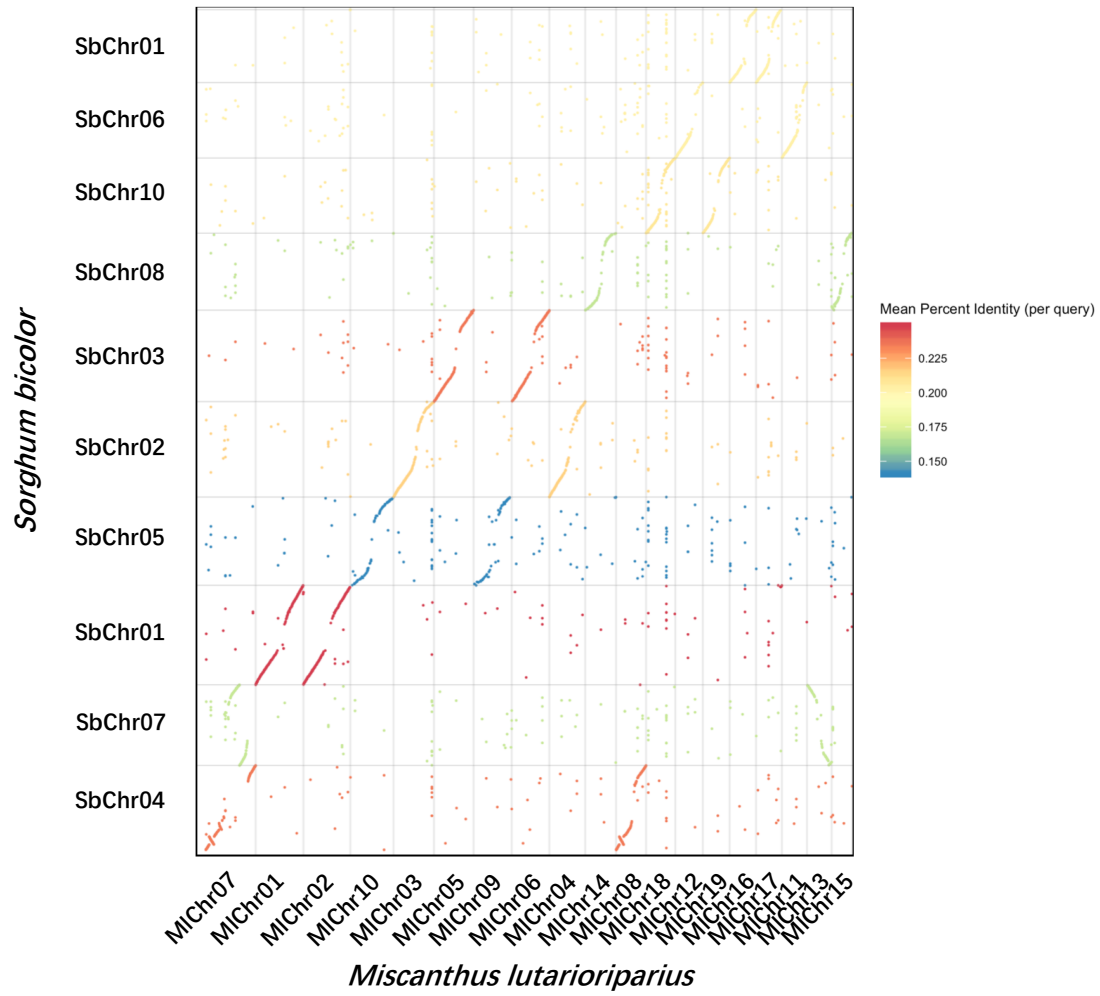
Supplementary Fig. 4 Interaction frequency distribution of Hi-C links among chromosomes.

The Hi-C version of *M. luteovoriparus* genome generated by 3d-dna pipeline.



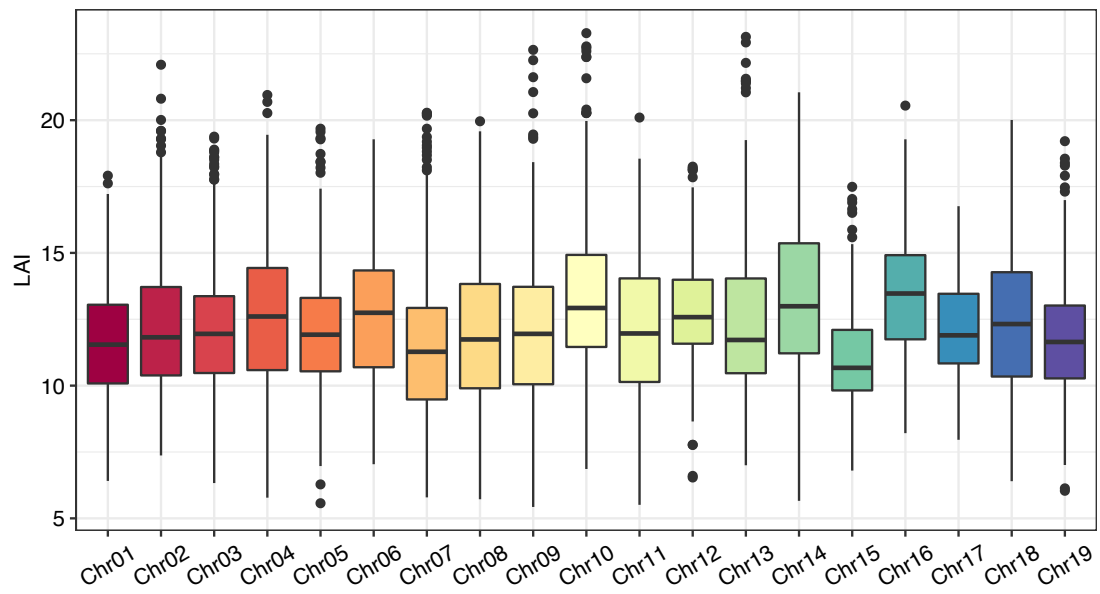
Supplementary Fig. 5 Comparison of sequences identity of chromosome anchoring results generated by LACHESIS and 3d-dna software.

The sequence comparison was performed using minmap2 (version 2.16) software.

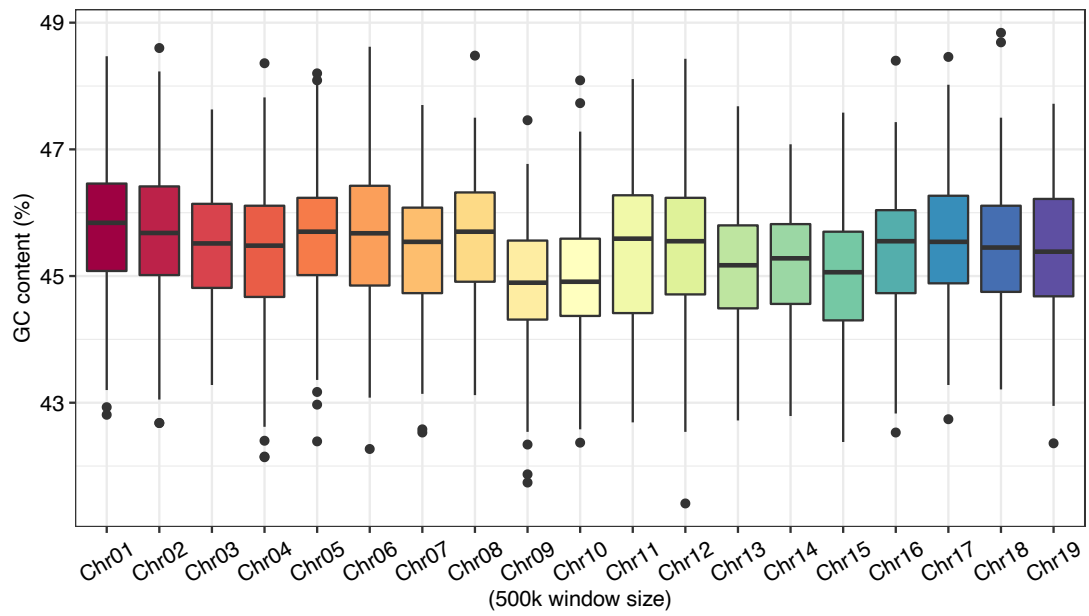


Supplementary Fig. 6 Comparison of whole genome sequence identity between *M. lutarioriparius* LACHESIS genome assembly and sorghum genome.

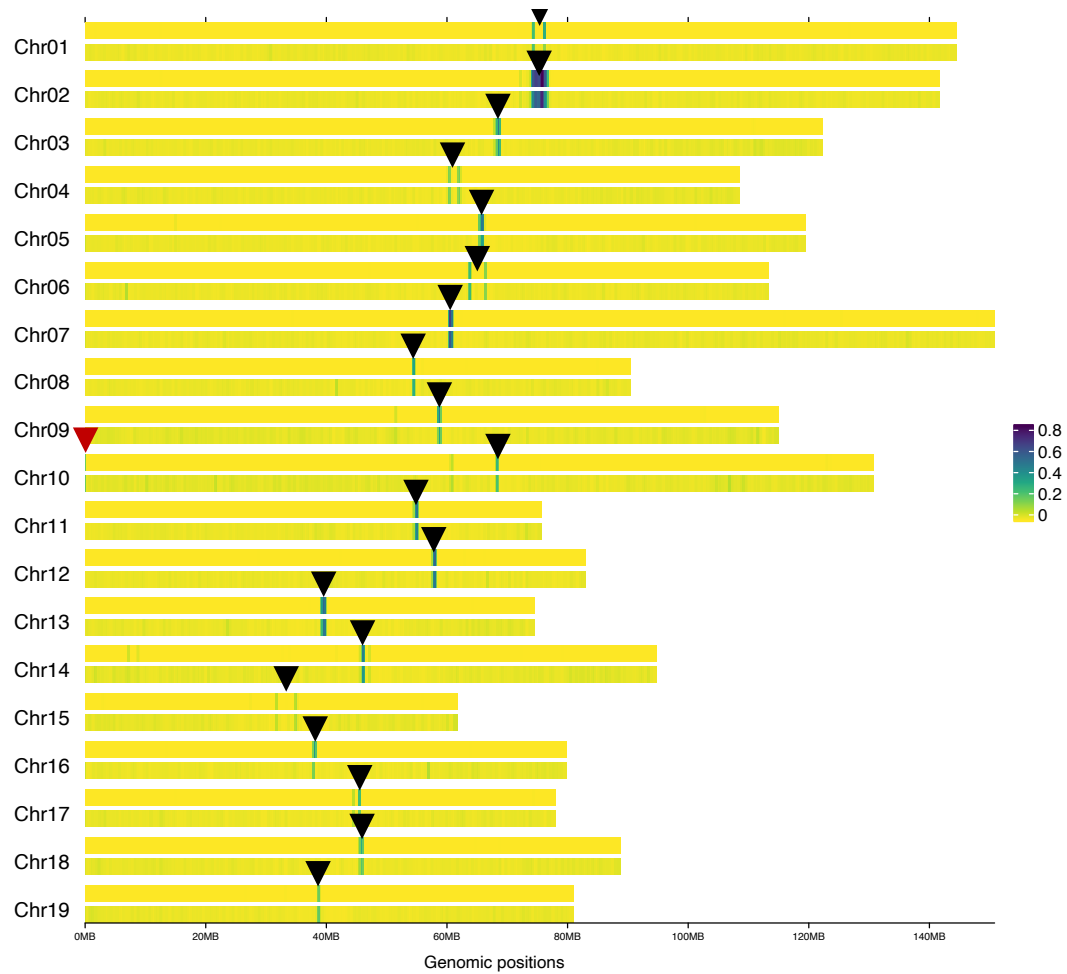
The sequence comparison was performed using minmap2 (version 2.16) software.



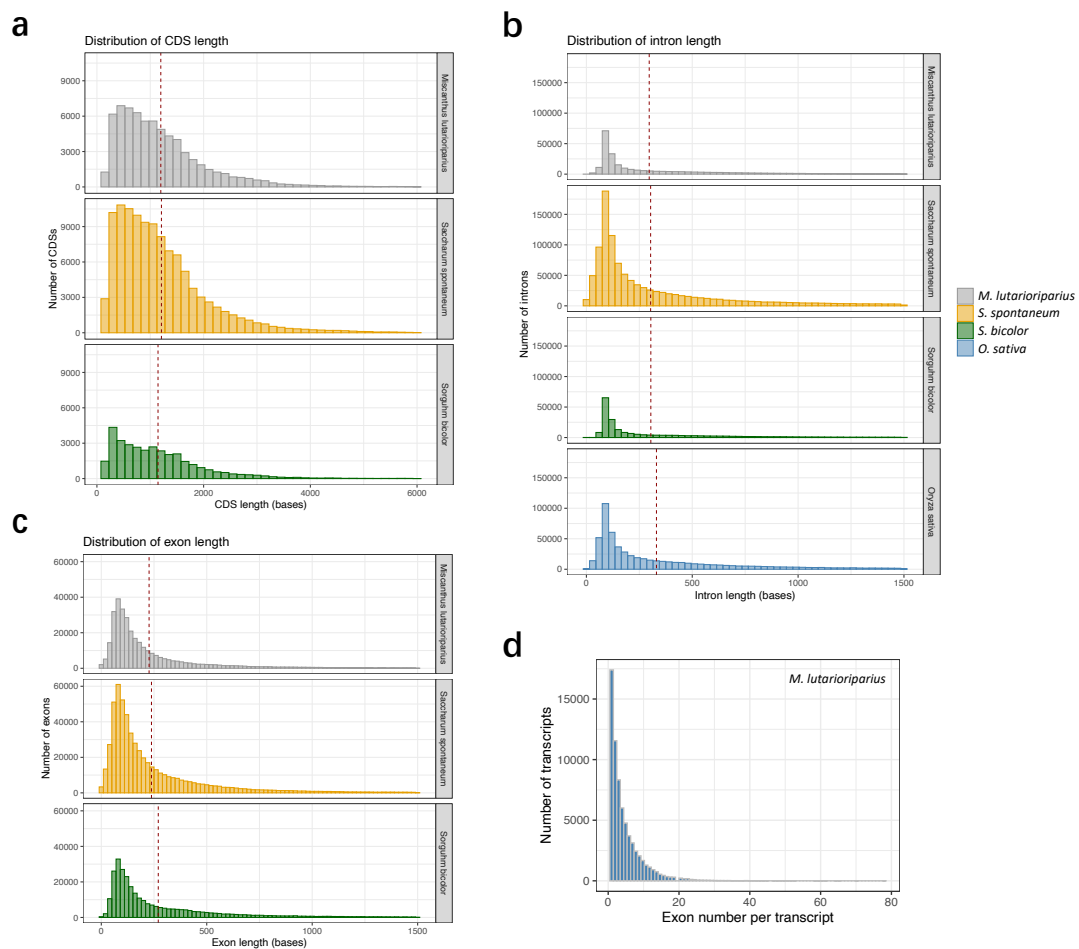
Supplementary Fig. 7 Boxplot of LTR Assembly Index (LAI) for 19 chromosomes.



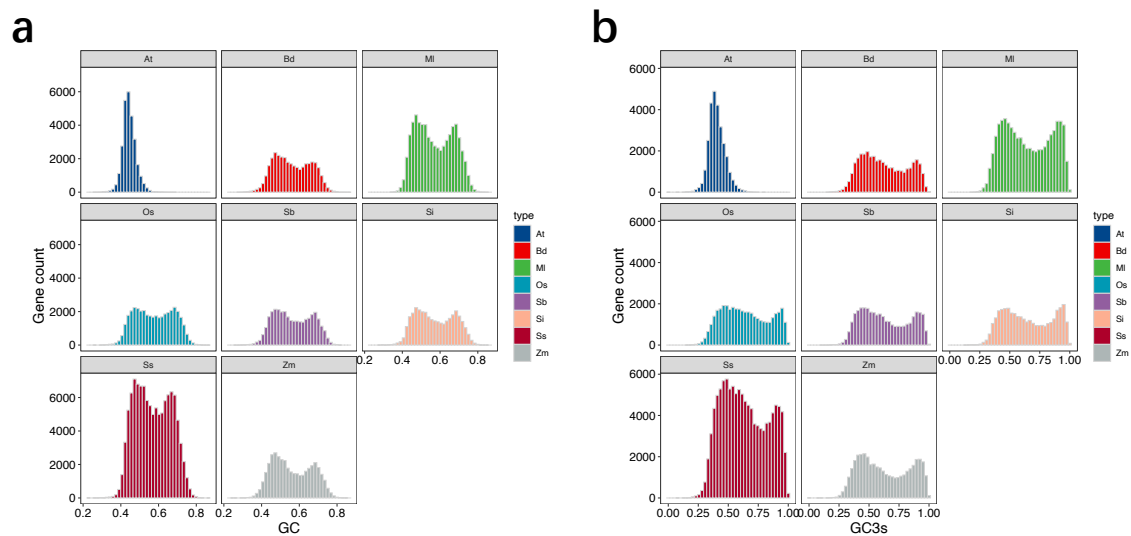
Supplementary Fig. 8 GC content of each chromosome of *M. lutarioriparius*.



Supplementary Fig. 9 Potential centromere and telomere position of *M. lutarioriparius* genome. The heatmap tracks for each chromosome represent the density of tandem repeats (non-overlap window size = 500 kb) identified by Tandem Repeat Finder (upper track) and the simple sequence repeats identified by RepeatMasker (lower track). Black and red inverted triangles were used to indicate the potential locations of centromere and telomere, respectively.

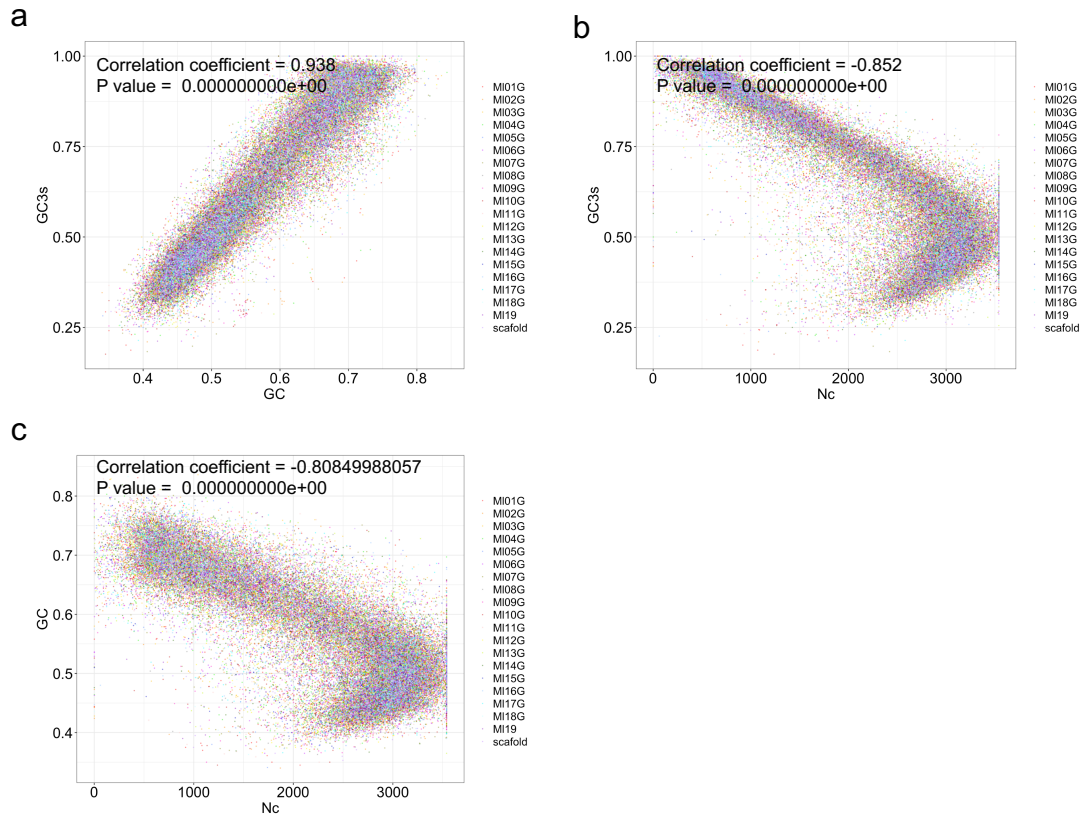


Supplementary Fig. 10 The length distribution of protein-coding sequence (CDS), exon and intron of *M. lutarioriparius*, *S. spontaneum* and *S. bicolor*.



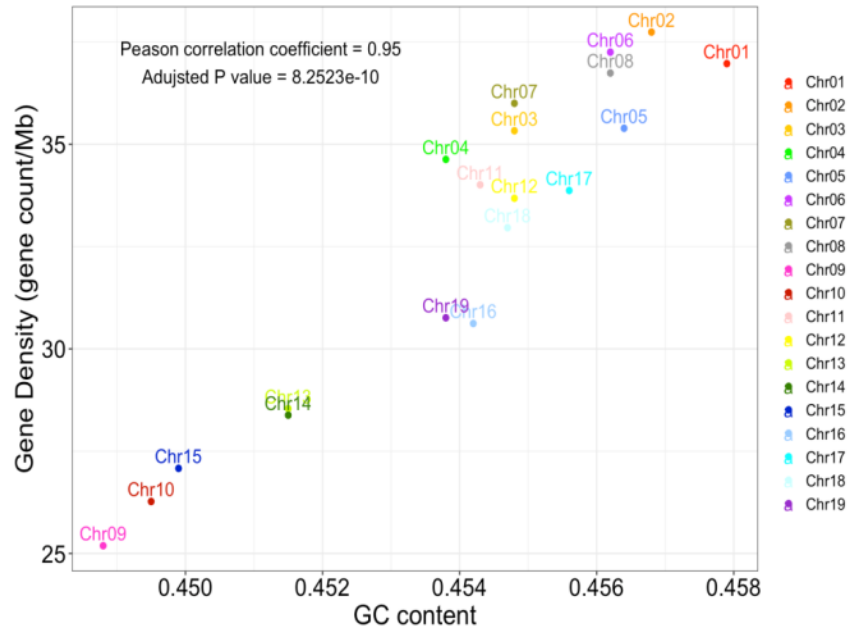
Supplementary Fig. 11 Comparison of GC content distribution (a) and GC3s content distribution (b) of whole genome coding sequences (CDS) for 8 species.

At: *Arabidopsis thaliana*, Bd: *Brachypodium distachyon*, Ml: *Miscanthus lutarioriparius*, Os: *Oryza sativa*, Sb: *Sorghum bicolor*, Si: *Seteria italica*, Ss: *Saccharum spontaneum*, Zm: *Zea mays*.

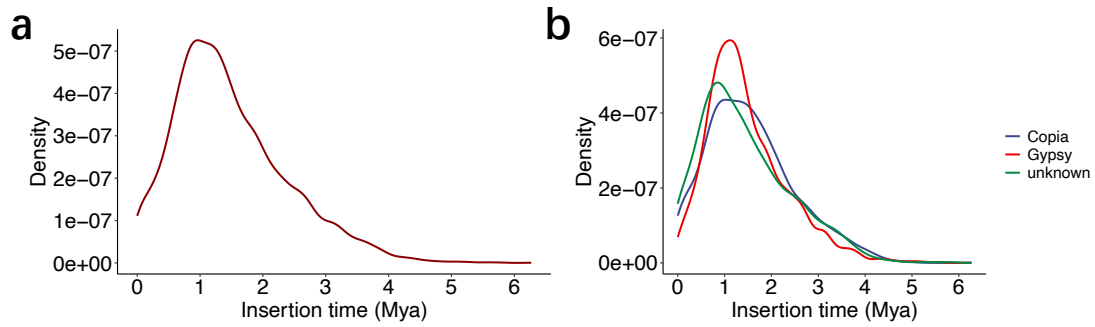


Supplementary Fig. 12 The correlation analysis of GC3s, GC content and effective number of codons (Nc).

The GC3s of *M. lutarioriparius* CDS has a strong positive correlation with GC content (a). The effective number of codons (Nc) of *M. lutarioriparius* CDS has negative correlation with GC and GC3s, respectively.

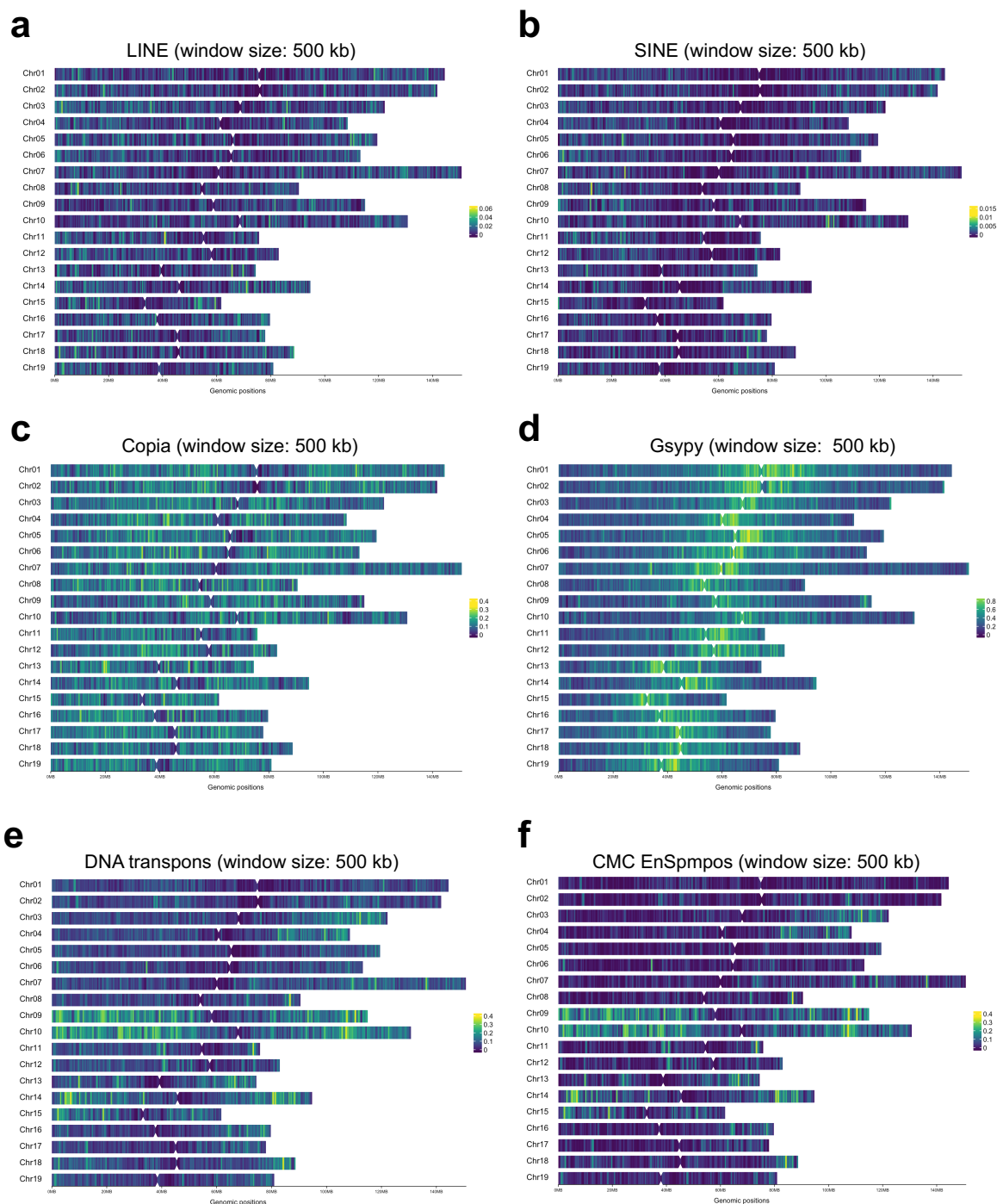


Supplementary Fig. 13 Correlation analysis of GC content and gene density.



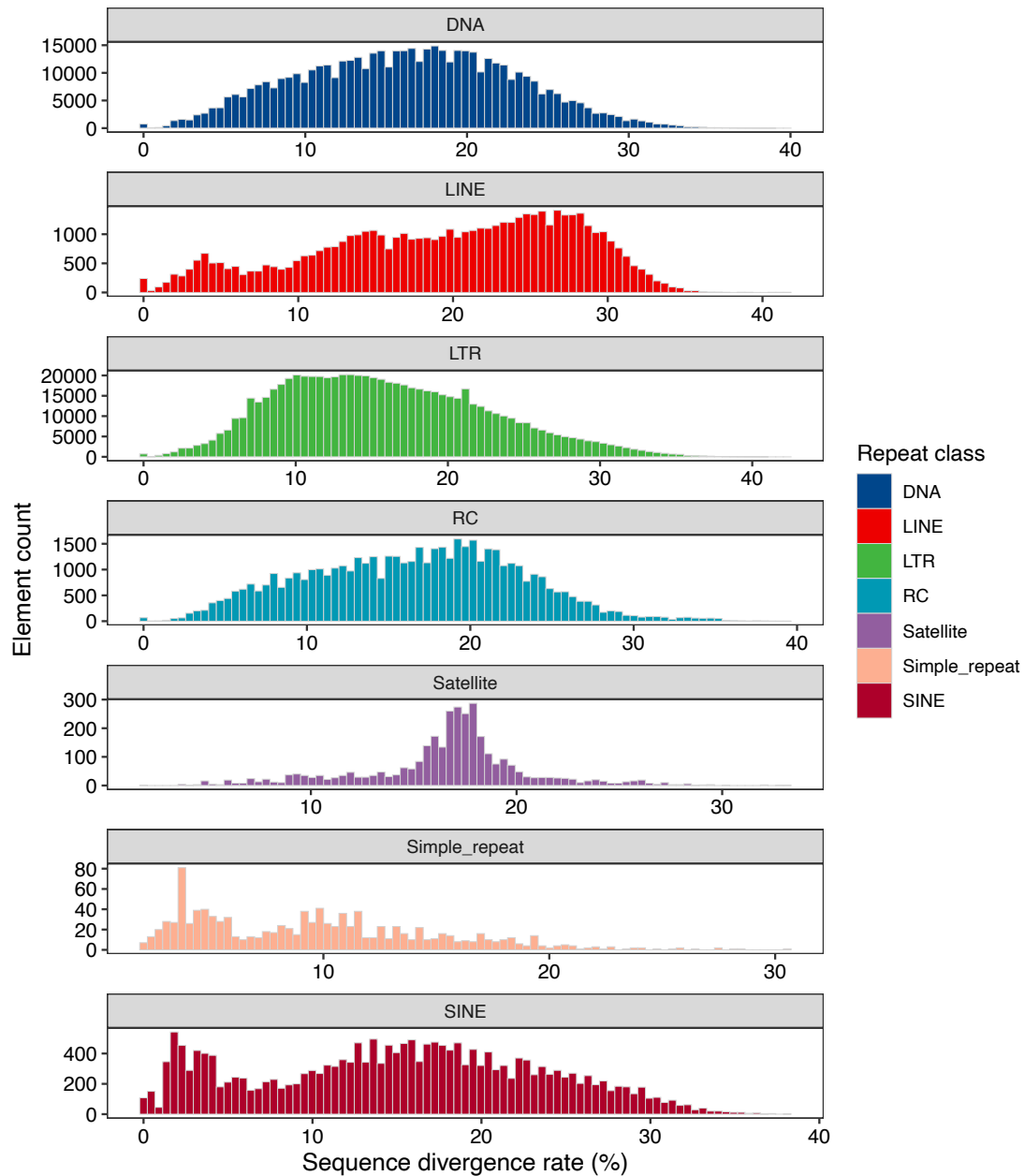
Supplementary Fig. 14 The frequency distribution of intact LTR-RTs insertion time.

- (a) The frequency distribution of insertion time of intact LTR-RTs in *M. lutarioriparius* genome.
- (b) The detailed frequency distribution of insertion time of intact *Copia*, *Gypsy* and other unclassified LTR-RTs in *M. lutarioriparius* genome.



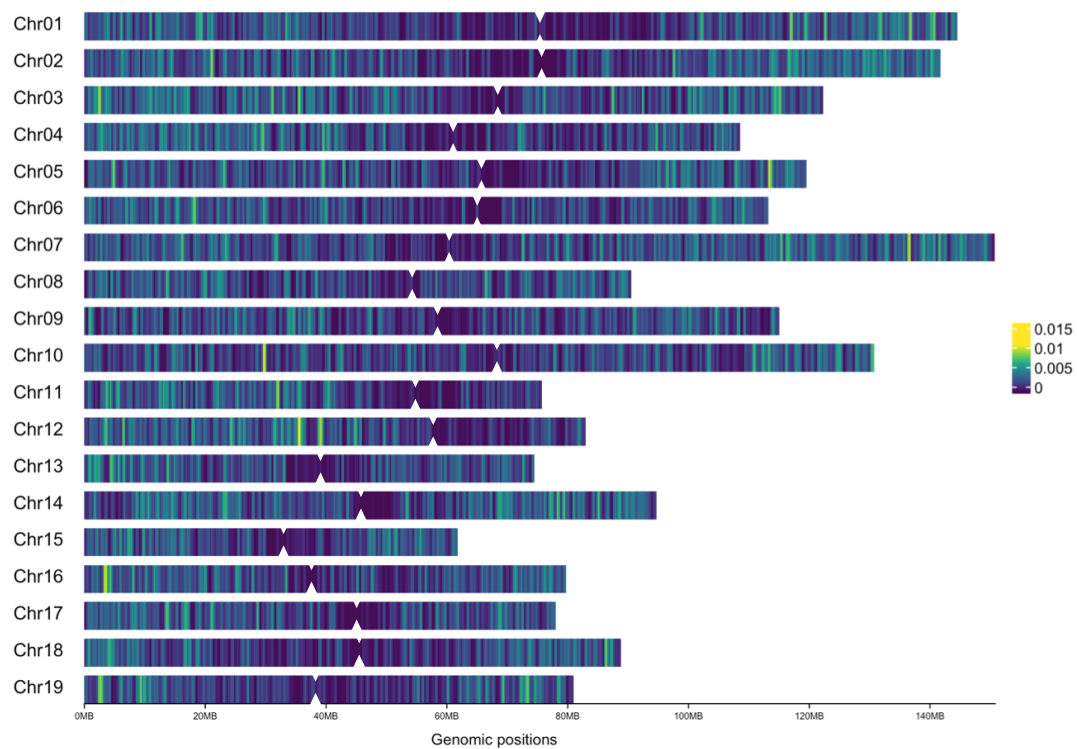
Supplementary Fig. 15 The density distribution of repeat elements across 19 chromosomes

The density for each type repeat elements was plotted using a non-overlap sliding window approach. The chromosomal incisions indicate the location of centromeres.



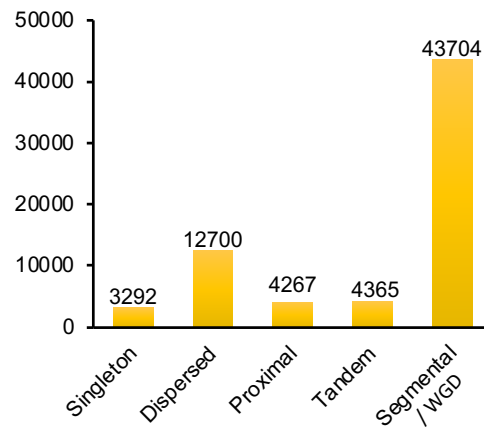
Supplementary Fig. 16 Distribution of divergence rates of different transposable element (TE) types in *M. lutarioriparius* genome.

Definitions of the abbreviations follow: DNA, DNA transposon; LINE, Long Interspersed Nuclear Element; LTR, Long Terminal Repeat retroelement; RC, RC/Helitron; SINE, Short Interspersed Nuclear Element.

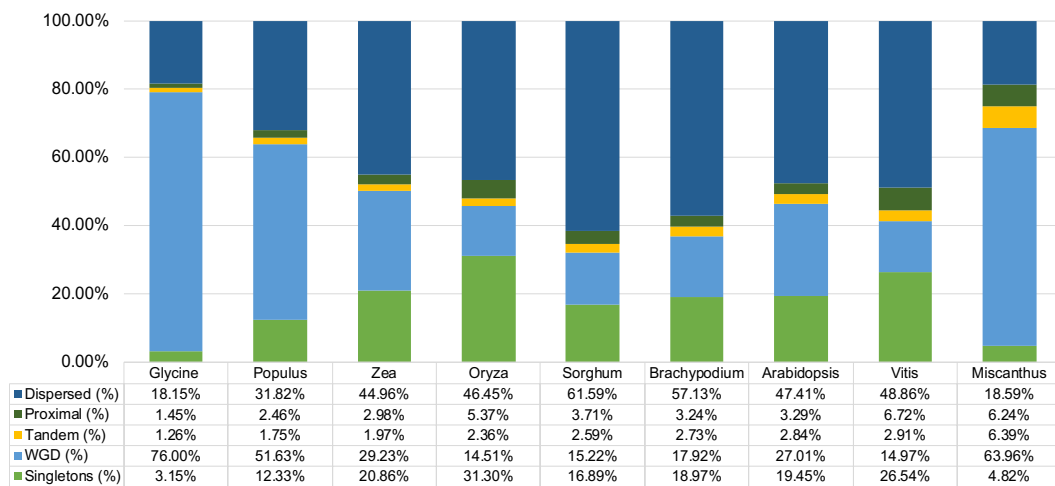


Supplementary Fig. 17 The density distribution of MITEs.

a

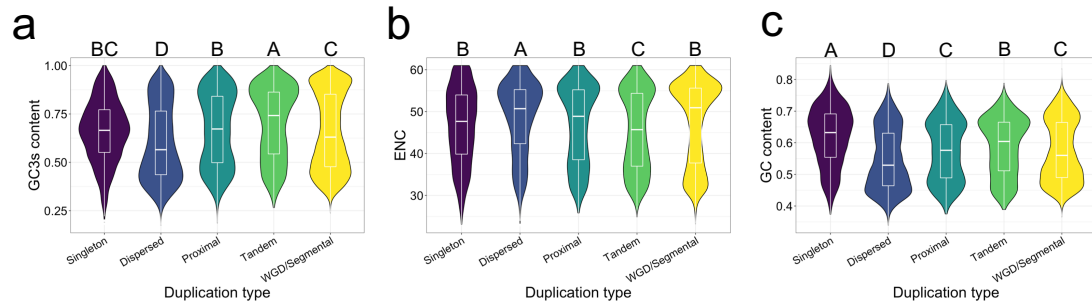


b



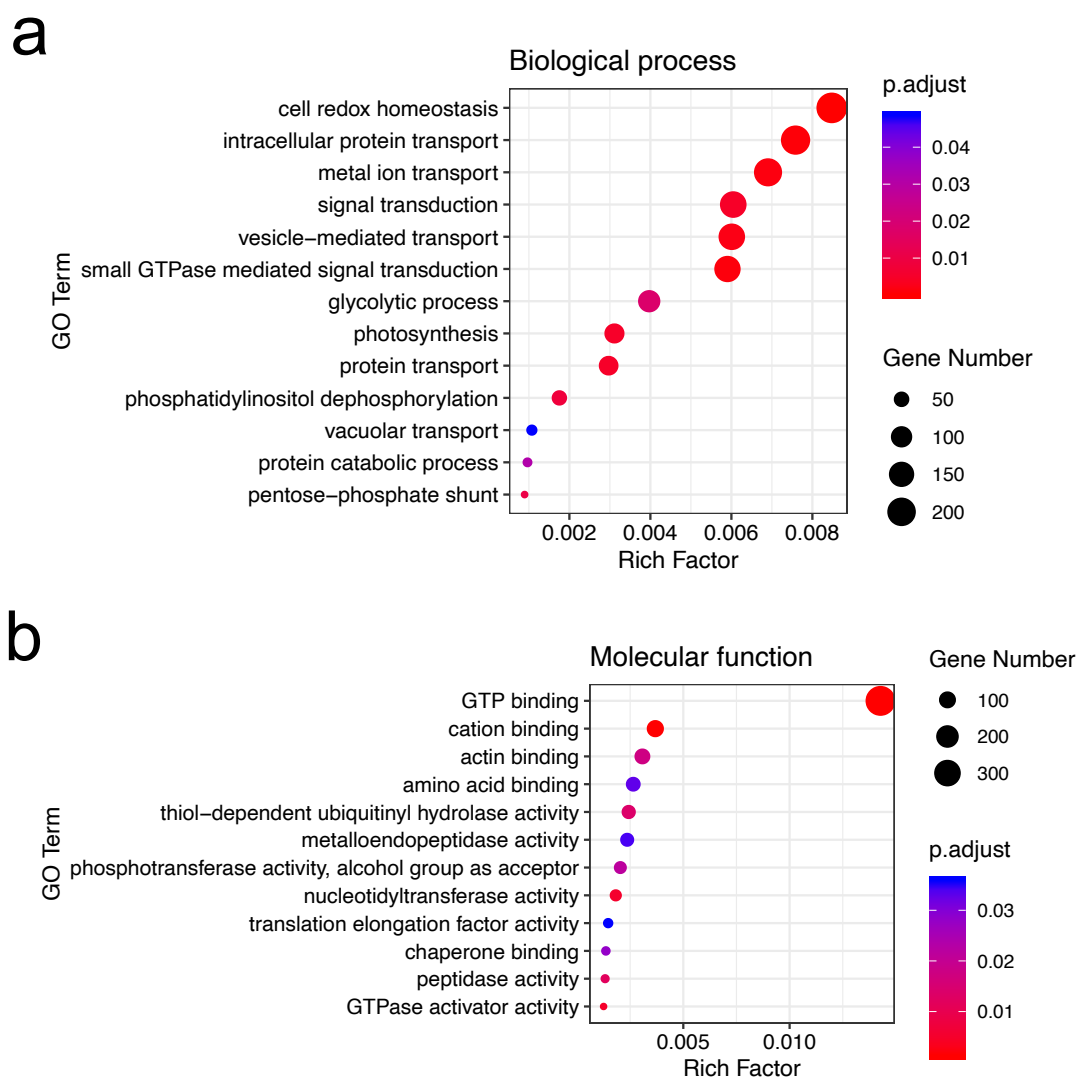
Supplementary Fig. 18 Different origins of genes in *M. lutarioriparius* genome.

(a) The number of genes for each duplication mode. The origin of genes in *M. lutarioriparius* were classified into singleton (no duplication), dispersed (duplication type other than WGD/segmental, tandem and proximal), proximal (two duplicated genes are distributed adjacent to each other on chromosomes, with no more than 10 genes spaced but not adjacent), tandem (consecutive repeat) and segmental/whole genome duplications (collinear genes in collinear blocks) using MCScanX (Wang et al., 2012). (b) The numbers of genes from different origins in nine angiosperm genomes. Except *Miscanthus*, the statistics of gene origins of other eight angiosperms come from the MCScanX software paper (Wang et al., 2012).



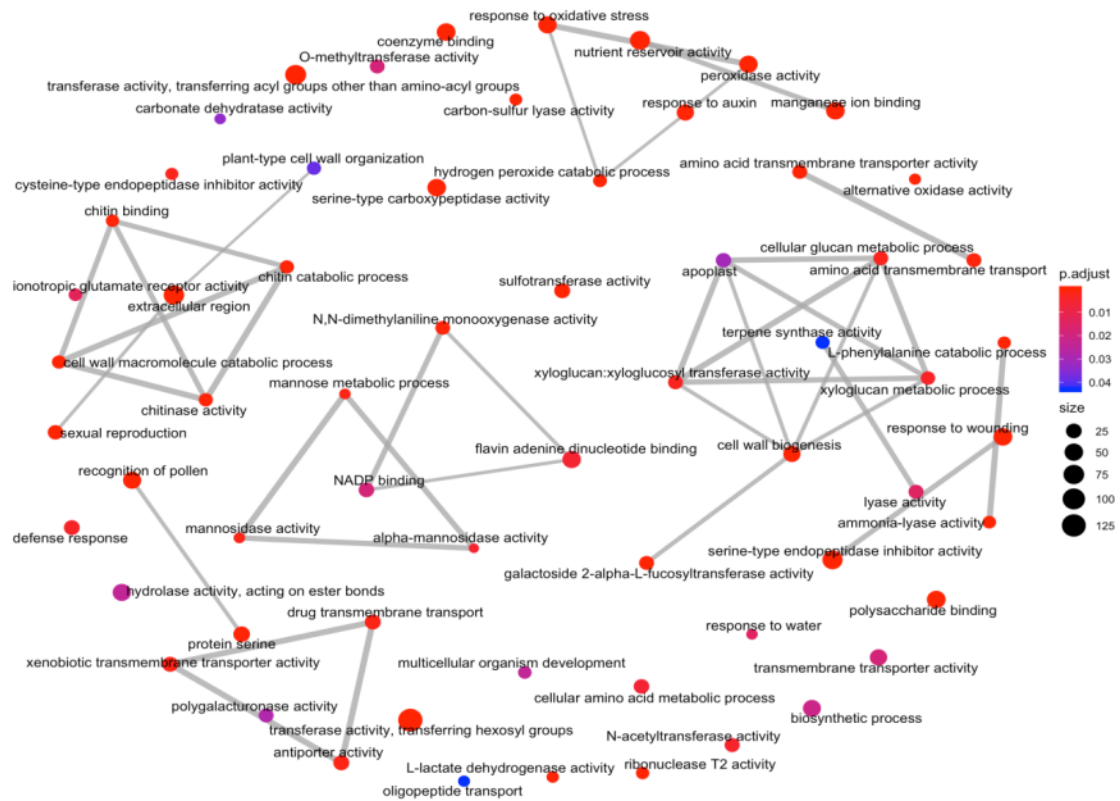
Supplementary Fig. 19 The distribution of GC3s, effective number of codons (ENC/Nc) and GC content for genes resulted from different duplication mode.

(a) GC3s content. (b) Effective number of codons (ENC/Nc). (c) GC content. The different duplication type (Singleton, Dispersed, Proximal, Tandem and WGD/Segmental) labeled with different capital letters differ significantly in GC3s, effective number of codons and GC content (Statistical comparison was carried out using ANVOA followed by Tukey Honest Significant Differences analysis, $P < 0.01$).



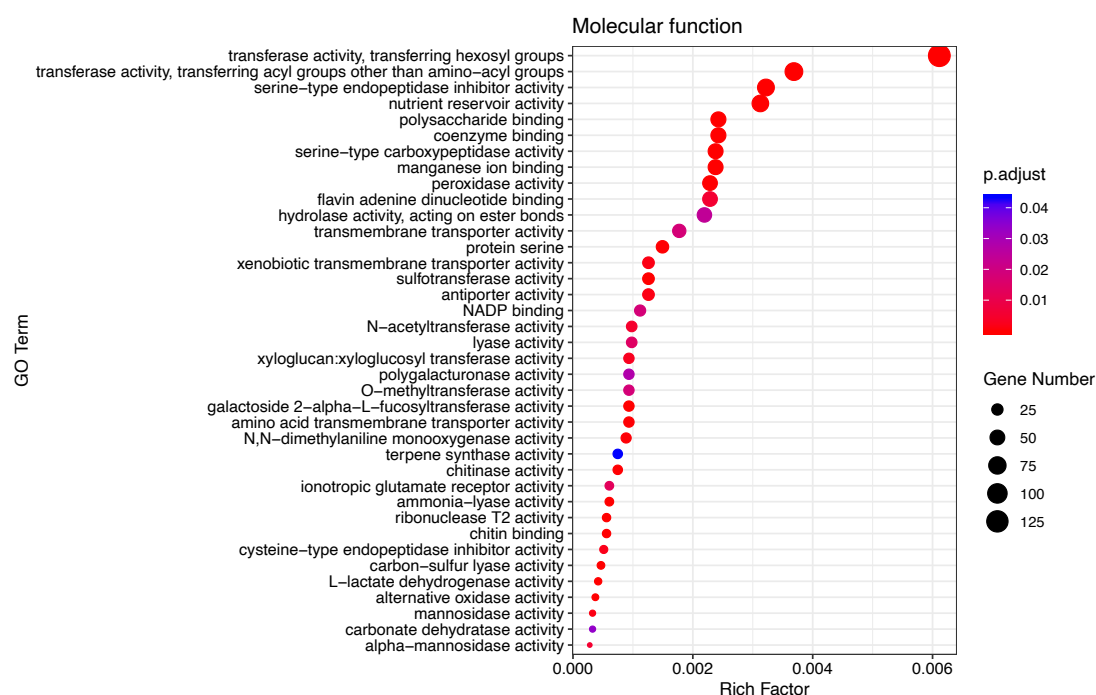
Supplementary Fig. 20 Gene Ontology (GO) enrichment analysis of *M. lutarioriparius* WGD/segmental duplication.

(a) GO enrichment results of biological process category. The color of circle represents the FDR (false discovery rate) in the hypergeometric test corrected using BH method (Benjamini and Hochberg, 1995). The size of circle represents the gene count of the GO terms. The rich factor is defined as gene count / total query gene count. (b) GO enrichment results of molecular function category.



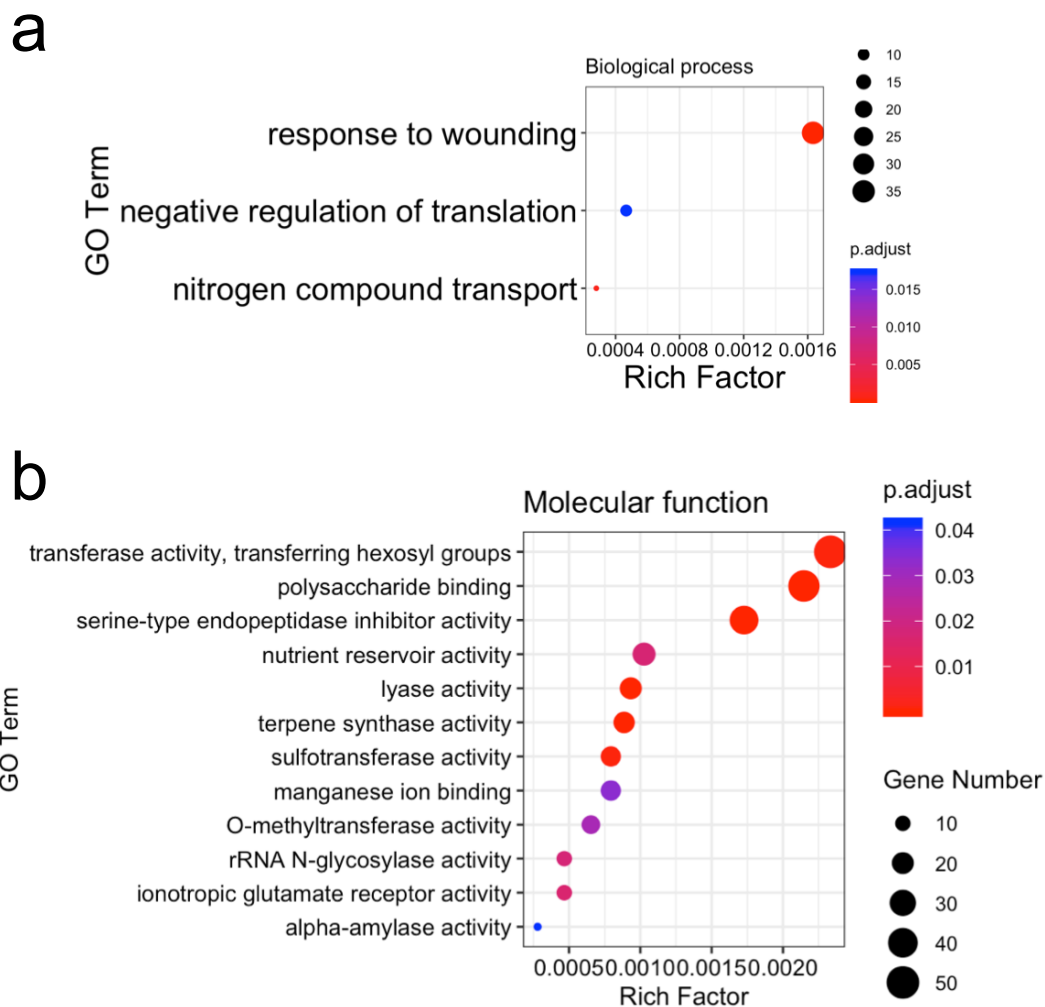
Supplementary Fig. 21 Enrichment Map for enrichment result of over-representation test for gene duplicates from tandem duplication.

Mutually overlapping gene sets tend to cluster together, making it easier for interpretation.



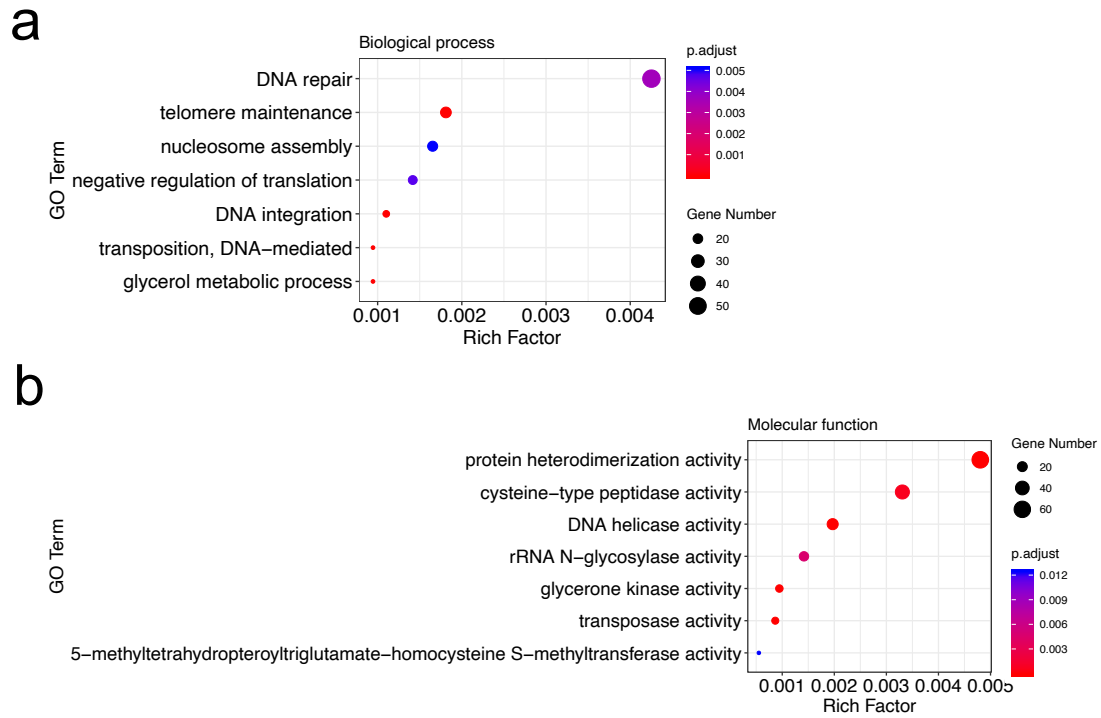
Supplementary Fig. 22 Gene Ontology (GO) enrichment analysis of *M. lutarioriparius* genes resulted from tandem duplication (molecular function category).

The color of circle represents the FDR (false discovery rate) in the hypergeometric test corrected using BH method (Benjamini and Hochberg, 1995). The size of circle represents the gene count of the GO terms. The rich factor is defined as gene count / total query gene count.



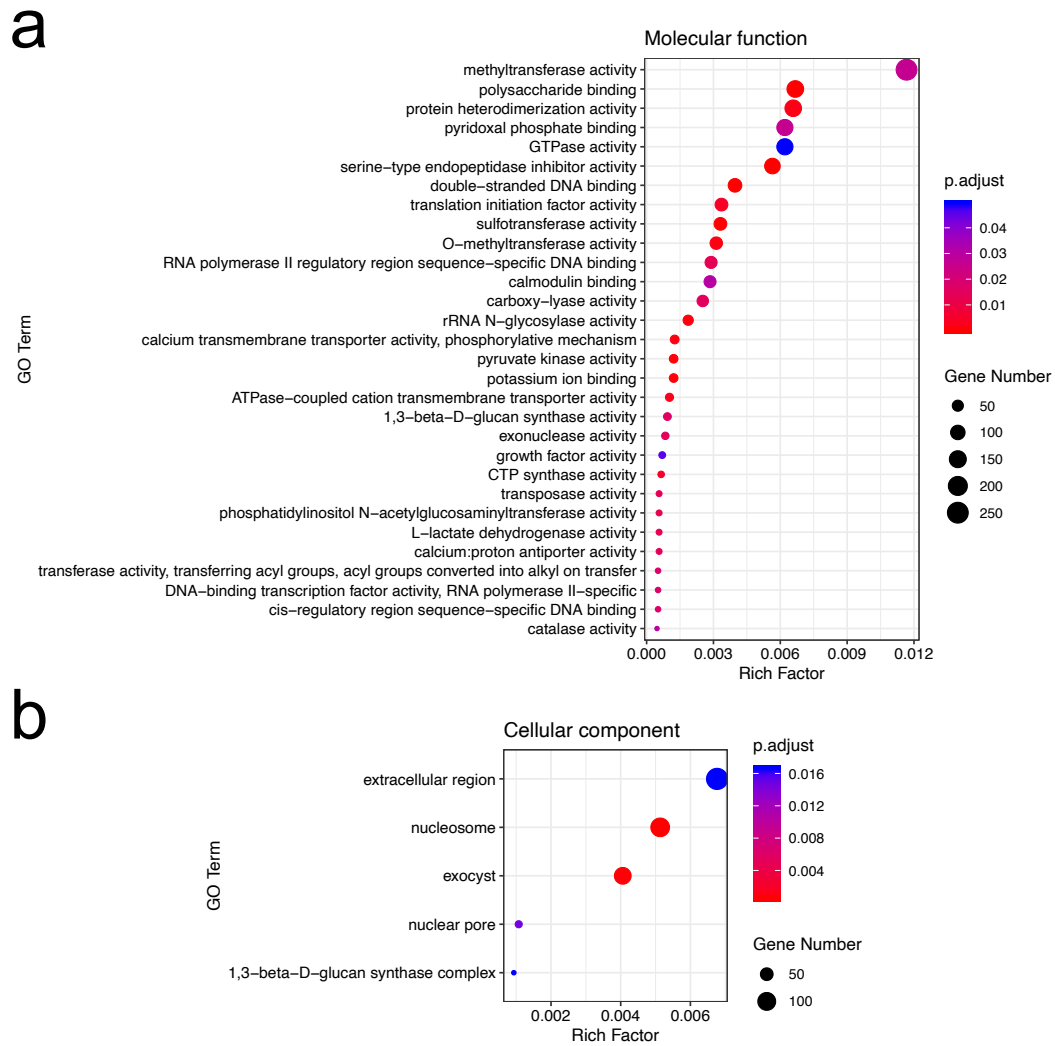
Supplementary Fig. 23 Gene Ontology (GO) enrichment analysis of *M. lutarioriparius* gene duplicates from proximal duplication.

The color of circle represents the FDR (false discovery rate) in the hypergeometric test corrected using BH method (Benjamini and Hochberg, 1995). The size of circle represents the gene count of the GO terms. The rich factor is defined as gene count / total query gene count.



Supplementary Fig. 24 Gene Ontology (GO) enrichment analysis of *M. lutarioriparius* gene duplicates of dispersed duplication origin

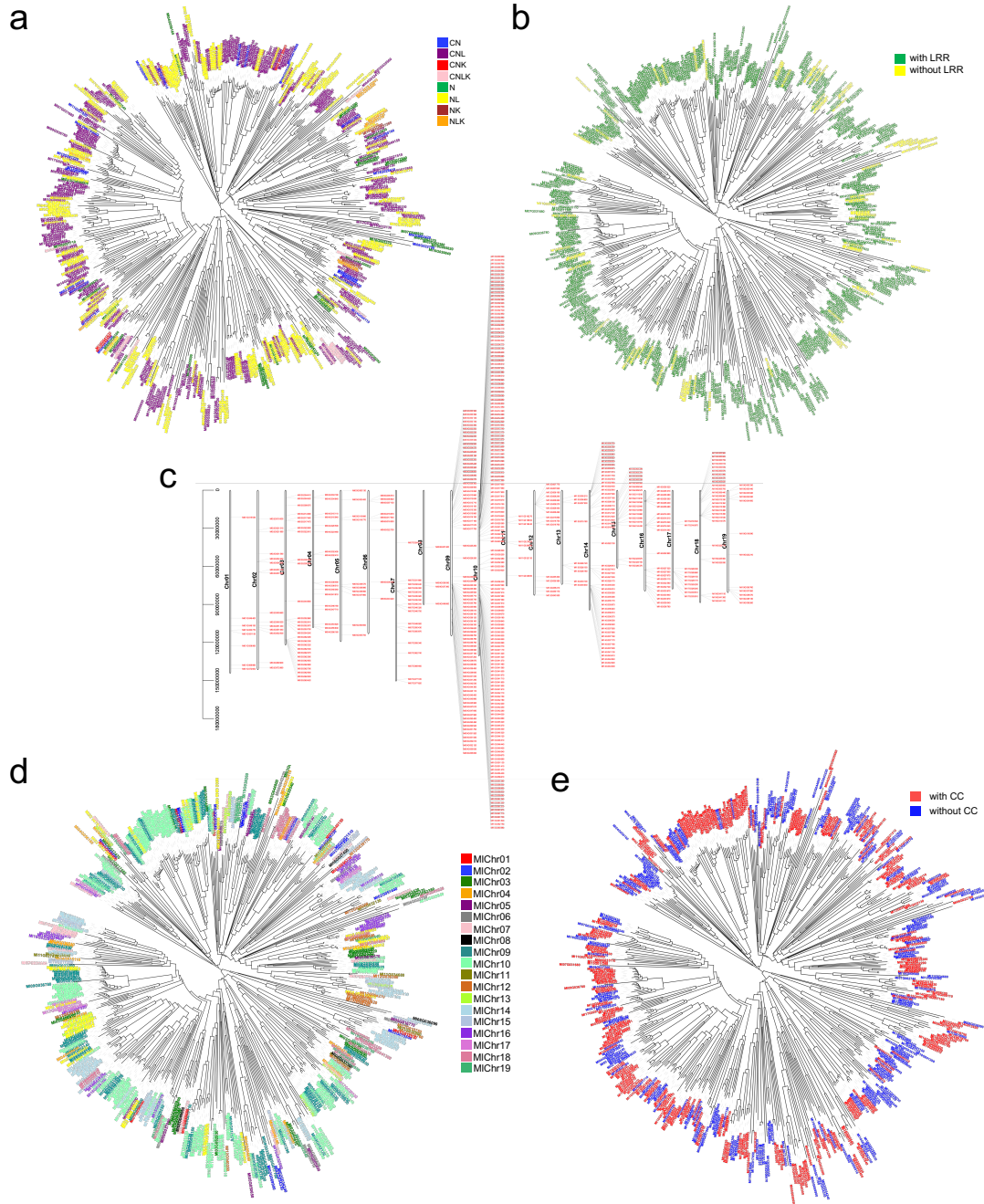
(a) GO enrichment results of biological process category. The color of circle represents the FDR (false discovery rate) in the hypergeometric test corrected using BH method (Benjamini and Hochberg, 1995). The size of circle represents the gene count of the GO terms. The rich factor is defined as gene count / total query gene count. (b) GO enrichment results of molecular function category.



Supplementary Fig. 25 Dot-plot of GO enrichment for 9,509 expanded gene families of *M.*

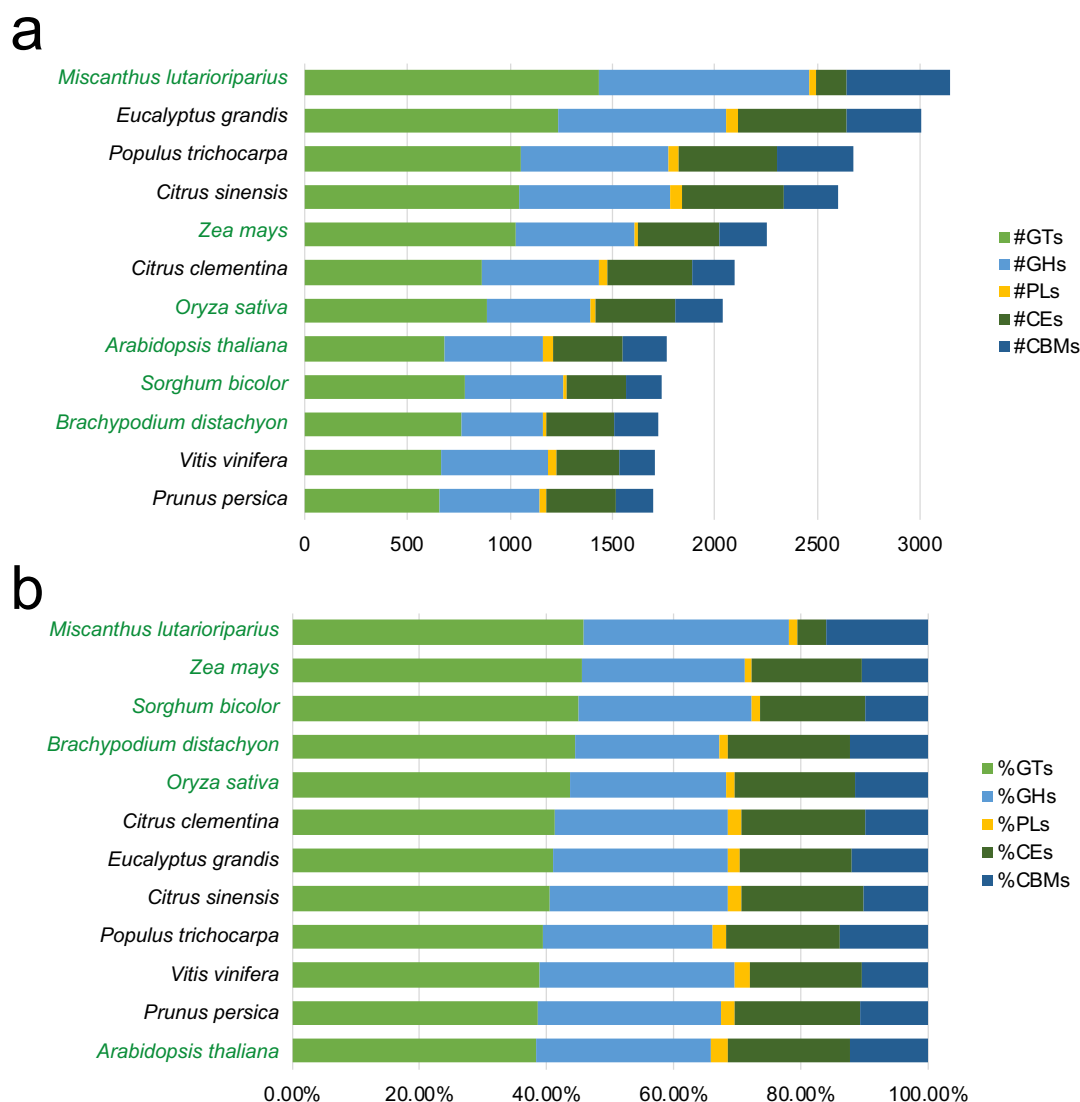
lutarioriparius

(a) GO enrichment results of biological process category. The color of circle represents the FDR (false discovery rate) in the hypergeometric test corrected using BH method (Benjamini and Hochberg, 1995). The size of circle represents the gene count of the GO terms. The rich factor is defined as gene count/total query gene count. (b) GO enrichment results of cellular component category.



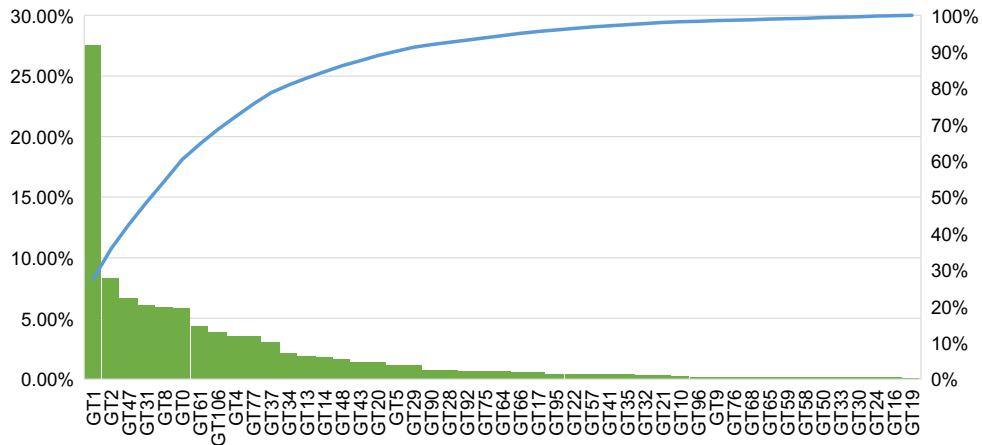
Supplementary Fig. 26 Phylogenetic tree reconstructed based on the amino acid sequences in the NBS domain of NBS-encoding genes from *M. luteoriparius*

(a) Color-coded by classification of NBS-encoding genes. (b) Color-coded by presence or absence of LRR domain. (c) Genomic distribution of NBS-encoding genes. (d) Color-coded by presence or absence of CC domain. (e) Color-coded by chromosome as indicated.

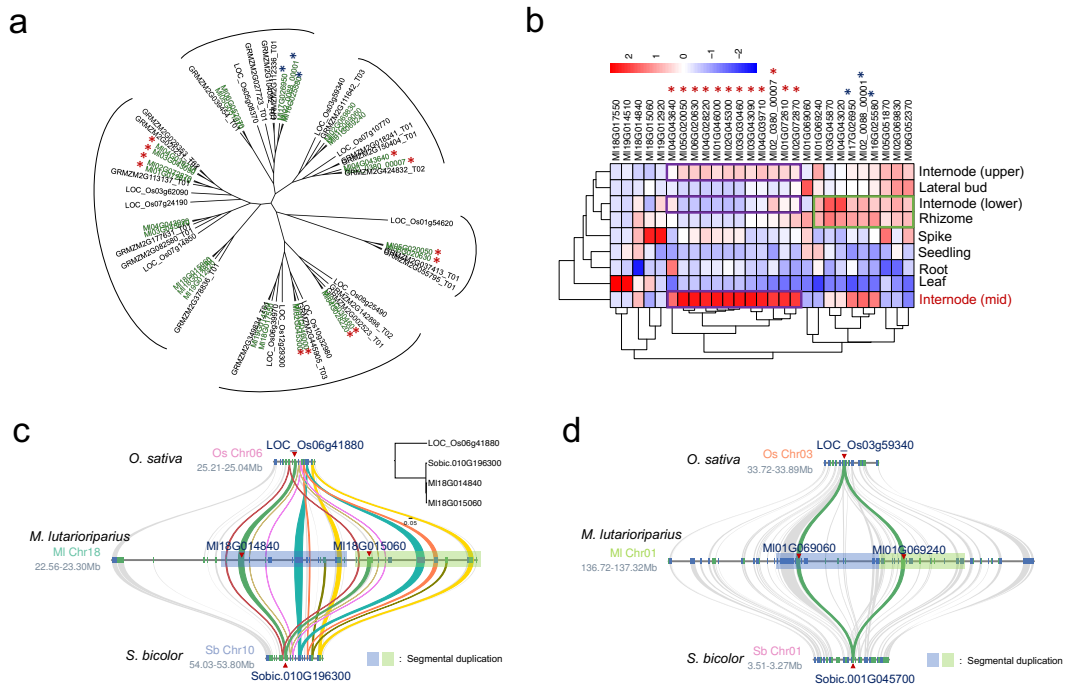


Supplementary Fig. 27 CAZyme domain classes frequency distribution in twelve plant species

(a) Absolute frequency of CAZyme domain classes in twelve plant species. Plant species are on the y-axis, and the frequency of CAZyme domains within all CAZyme genes is shown on the x-axis. The green color of plant species is used to indicate the herbaceous species, while the black is used to indicate the woody species. The glycosyl transferase (GT) domain class is indicated using green, glycosyl hydrolase (GH) domain class using blue, polysaccharide lyase (PL) domain class using yellow, carbohydrate esterase (CE) domain class using dark green and carbohydrate binding module (CBM) domain class using dark blue. Except *Miscanthus lutarioriparius*, the other data came from Pinard et al. (Pinard *et al.*, 2015). (b) Relative frequency of CAZyme domain classes in twelve plant species. The relative frequency of carbohydrate active enzyme (CAZyme) domain classes, as a percentage, is shown on the x-axis. The species of plant is shown on the y-axis.

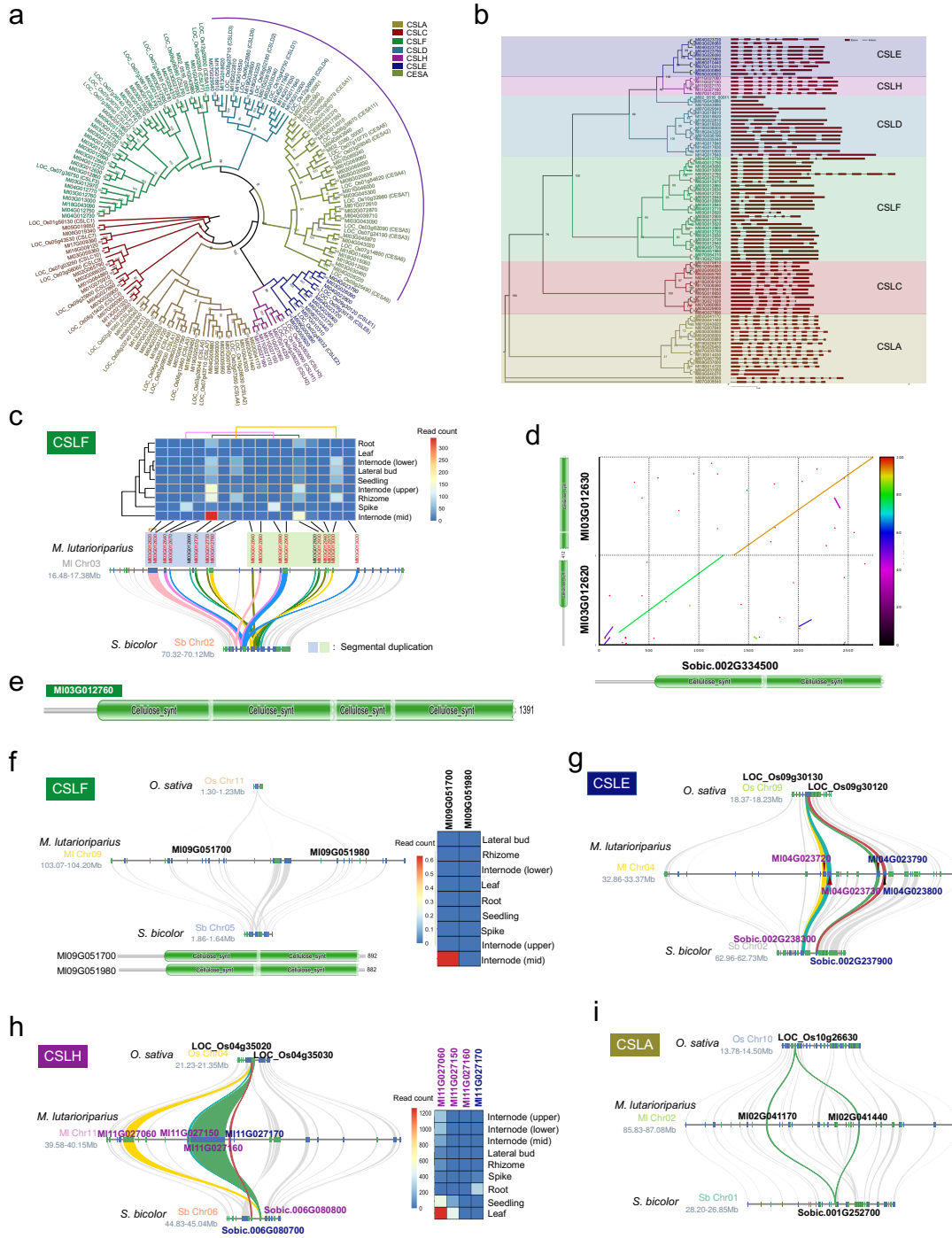


Supplementary Fig. 28 Frequency distribution of gene members of GT family



Supplementary Fig. 29 Cesa genes of *M. lutarioriparius*

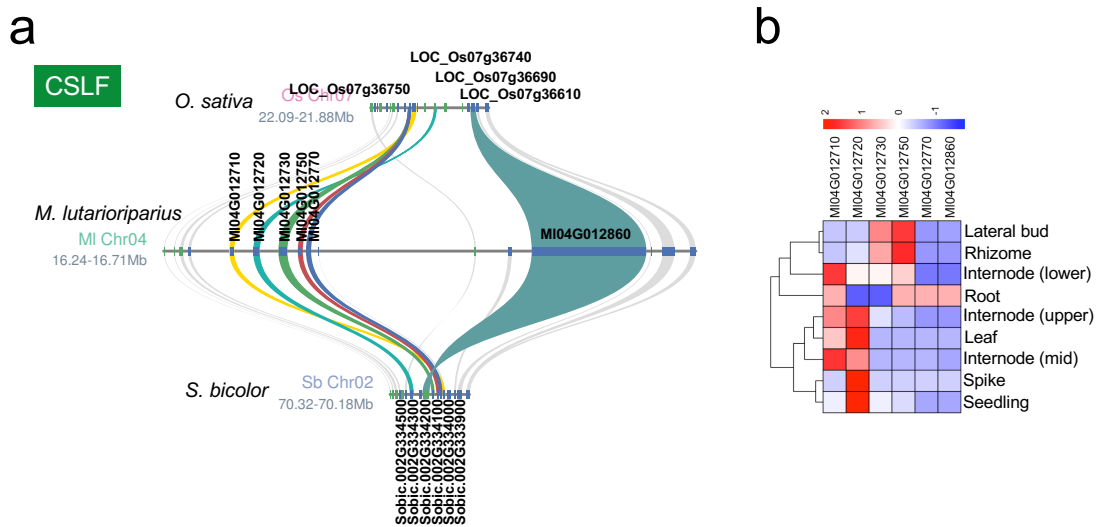
(a) Phylogeny tree of Cesa genes reconstructed using maximum likelihood approach by running IQtree software using protein sequences of *M. lutarioriparius*, rice and maize. Green color is used to indicate the genes of *M. lutarioriparius*. (b) Expression pattern of Cesa genes in 9 transcriptome samples. Red and blue stars were used to indicate the Cesa genes that have specifically high expression in the mid of internode. (c) Segmental duplication leads to the expansion of *M. lutarioriparius* Cesa genes (M18G014840 and M18G015060). (d) Segmental duplication leads to the expansion of *M. lutarioriparius* Cesa genes (M01G069060 and M01G069240)



Supplementary Fig. 30 Csl gene family of *M. lutarioriparius*

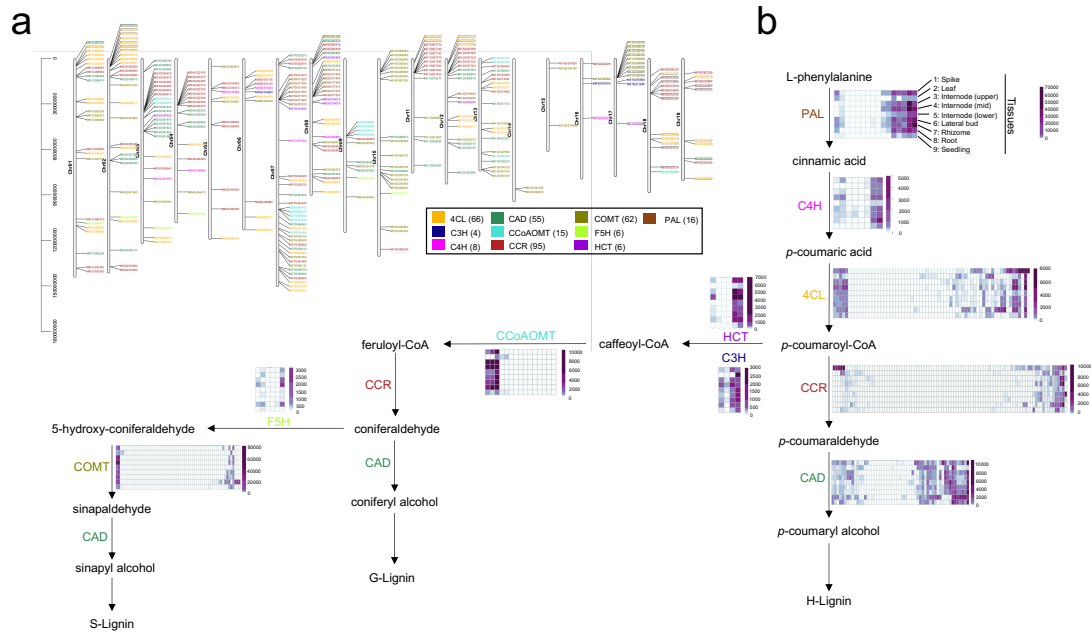
(a) Phylogeny tree of Csl and CesA of *M. lutarioripaiurs* was reconstructed using maximum likelihood approach. Different colors were used to indicate the groups. Bootstrap values (integer) are indicated at the nodes. (b) Gene structure diagram of Csl gene family of *M. lutarioriparius*. Red boxes and black lines were used to indicate the exons and introns, respectively. (c) The segmental duplication of CslF genes was occurred in *M. lutarioriparius* after it's split with sorghum. And expression heatmap of CslF gene family in 9 transcriptome samples. Links with same color is used to indicate the syntenic gene pairs that share common sorghum genes. The red gene symbols of *M.*

lutariopariarius are CslF genes. The expression pattern of CslF genes in 9 transcriptome samples are shown using heatmap. (d) Comparison of sequence similarity between MI03G012620 and MI03G012630. (e) Gene fusion occurred in MI03G012760. Pfam domain of Cellulose_synt (PF03552) is indicated by green box. (f) *M. lutariopariarius* specific duplication of CslF genes, which have no homolog in the expected collinear regions in both sorghum and rice genomes. These two Csl genes were confirmed by Pfam annotation. Heatmap is used to show the library-size normalized read count of MI09G05170 and MI09G051980, showing very low expression for both genes in 9 transcriptome samples. (g) Tandem duplication of CslE genes occurred in *M. lutariopariarius* after its split with sorghum. (h) Tandem and proximal duplication of CslH genes occurred in *M. lutariopariarius* after its split with sorghum. (i) Proximal duplication of CslA genes in *M. lutariopariarius* after its split with sorghum. The green lines were used to indicate the CslA genes collinearity between *M. lutariopariarius* and sorghum.



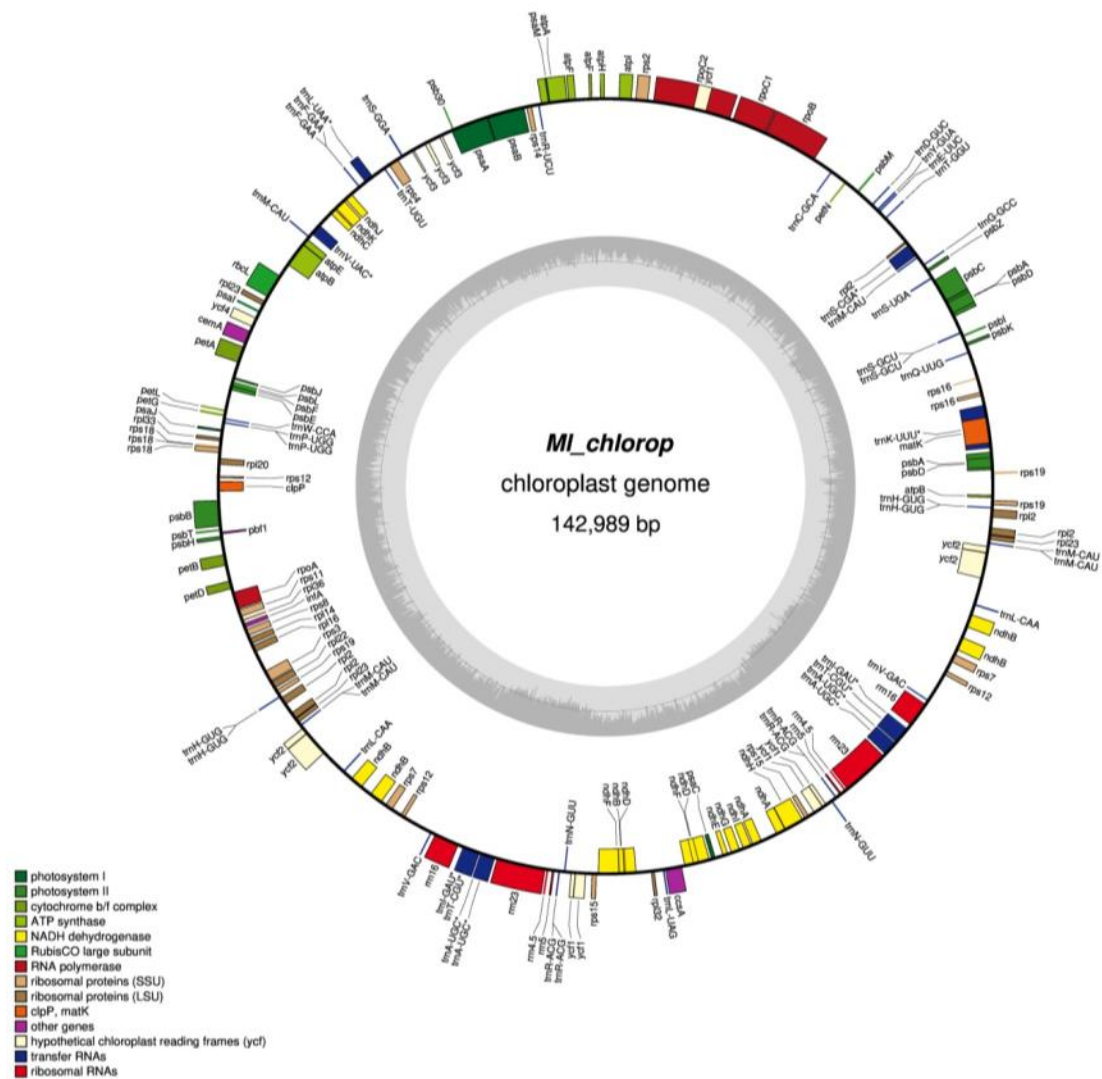
Supplementary Fig. 31 The marcosynteny of CslF family among *M. lutariopariarius*, rice and sorghum

(a) The syntenic gene pairs between *M. lutariopariarius*, rice and sorghum. (b) The expression pattern of gene members in CslF family of *M. lutariopariarius*.

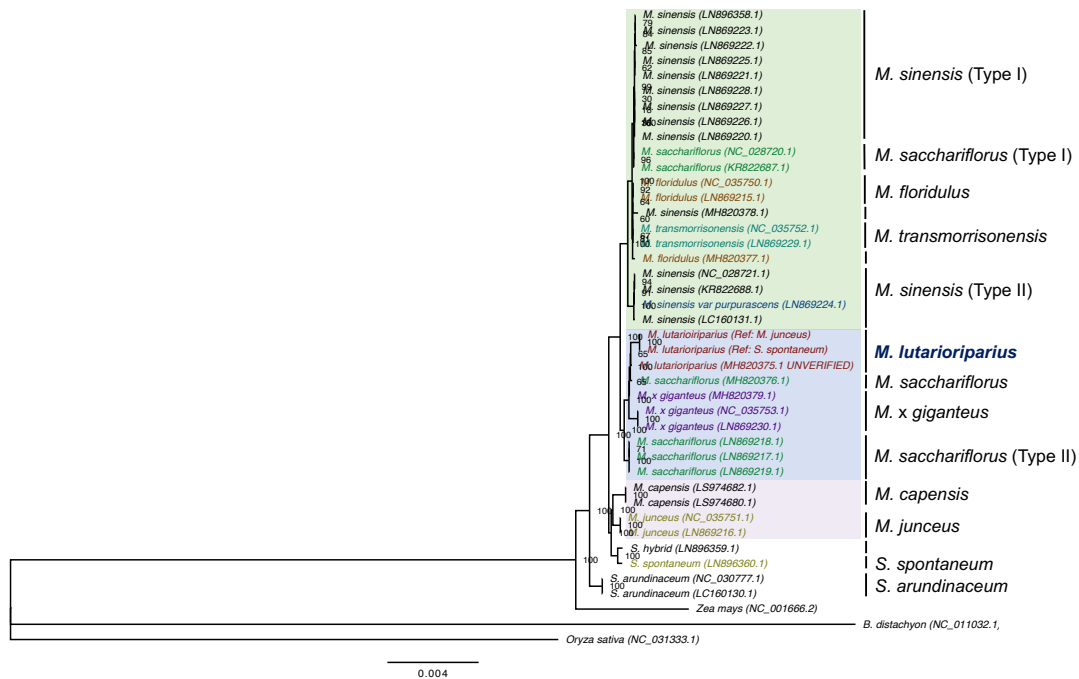


Supplementary Fig. 32 Genes involved in lignin biosynthesis in *M. lutarioriparius*

4CL: 4-coumarate: CoA ligase; C3H: p-coumarate-3-hydroxylase; C4H: Trans-cinnamate 4-hydroxylase; CAD: cinnamyl alcohol dehydrogenase; CCoAOMT: Caffeoyl coenzyme A-3-O-methyltransferase; CCR: cinnamoyl CoA reductase; COMT: caffeic acid O-methyltransferase; F5H: ferulate-5-hydroxylase; HCT: hydroxycinnamoyl-CoA shikimate/quinic hydroxycinnamoyl transferase; PAL: phenylalanine ammonia-lyase. The numbers in parentheses indicate the number of genes.



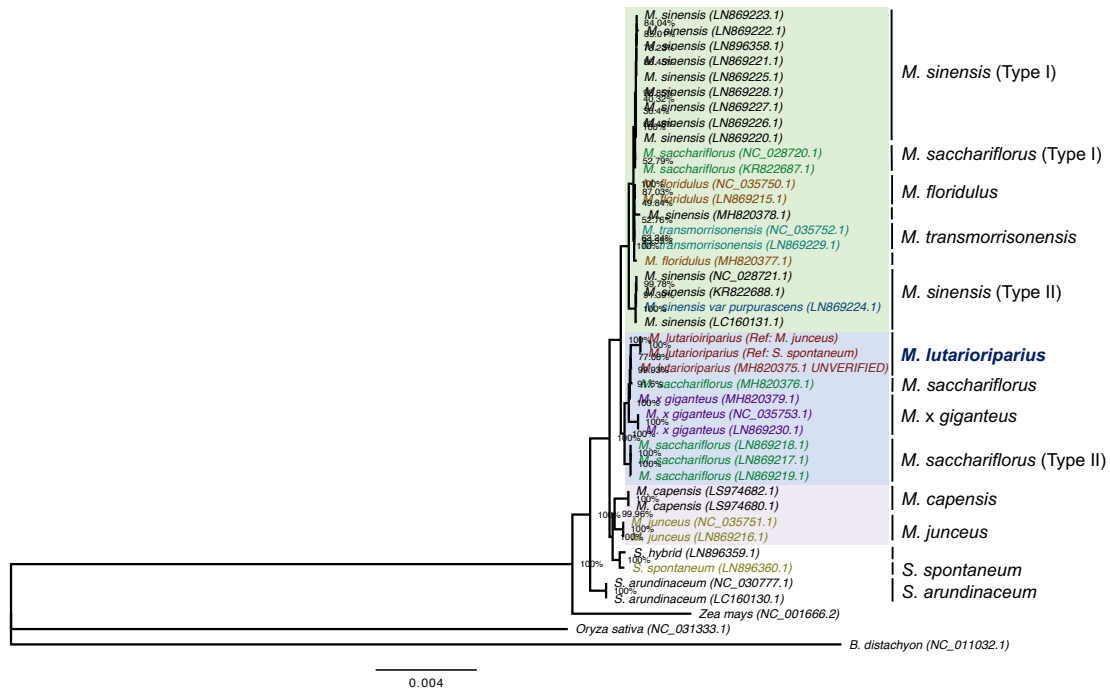
Supplementary Fig. 33 Visualization of annotation of *M. lutarioriparius* chloroplast genome assembled in this study



Supplementary Fig. 34 Phylogenetic tree of genus *Miscanthus* reconstructed using Maximum

Likelihood (ML) approach based on whole chloroplast genome sequence.

The chloroplast genome sequences were aligned using MAFFT (Kato and Standley, 2013) and the conserved sequence blocks were extracted using Gblock. The ML tree was reconstructed using IQ-TREE software with automatic best-fit model selection. The values of bootstrap support from 1000 bootstrap replicates using IQ-TREE ultrafast bootstrap algorithm were indicated at nodes.



Supplementary Fig. 35 Phylogenetic tree of genus *Miscanthus* reconstructed using Neighbor-Joining (NJ) method based on whole chloroplast genome sequence.

The chloroplast genome sequences were aligned using MAFFT (Kato and Standley, 2013) and the conserved sequence blocks were extracted using Gblock. The NJ tree was reconstructed using MEGA X (Kumar *et al.*, 2018) and 1000 replicates bootstrap test were carried out. The values of bootstrap were indicated at nodes.