First Chromosome-scale Genome Assembly of *Arundo donax*

Authors & Affiliations

Mengmeng Ren[1,+,*], Fupeng Liu[1], Xiaohong Han[1]

[1]Wuhan Rundo Biotechnology Co., Ltd. Wuhan, Hubei Province, 430300, China

[+,*]First author and corresponding author: Mengmeng Ren (ren_mm1991@163.com)

Abstract

Arundo donax L (*A. donax*). is a promising energy crop with high biomass and wide adaptability, while lack of reference genome limiting the genetic improvement of this crop. Here, we report two chromosome-scale assemblies of *A. donax* genome (0004 and 0408) using Pacbio SMRT sequencing and Hi-C technology. The genome size of these two assemblies are 1.30 Gb / 2.86 Gb with contig N50 33.15 Mb / 34.62 Mb respectively. A total of 74,403 / 167,586 gene models were predicted in 0004 and 0408 genome, of which over 90% genes were functionally annotated. Our result will enhance genetic understanding and promote the genetic improvement of *A. donax*.

Keywords: *A. donax*, genome assembly, genome annotation

1. Introduction

Arundo donax L. (*A. donax*) is a tall perennial grass that is commonly known as giant reed. *A. donax* is a widely utilized plant species with various applications in bioenergy, papermaking, and even as an ornamental lawn grass due to its fast growth and high biomass production [1,2,3,4]. One of the significant characteristics of *A. donax* is its high productivity and fast growth rate. The plant matures rapidly and is capable of producing large quantities of biomass in a short time, which makes it an attractive crop for bioenergy production [5,6,7]. Additionally, *A. donax* is highly tolerant of a broad range of environmental conditions, including drought, excess water, and high salinity [8,9]. These attributes make the plant an excellent candidate for sustainable agriculture and bioenergy development.

A. donax is also known for its environmental and ecological benefits. It is considered to be an excellent tool for phytoremediation, which could be used to clean up contaminated soil and water [10,11]. *A. donax* is effective in removing heavy metals and other pollutants from contaminated sites, making it a valuable resource in environmental restoration [12]. In recent years, there has been increased interest in the study and utilization of *A. donax*, which has led to a number of research efforts aimed at understanding its biological properties.

The sterile nature of *A. donax* limits the genetic variability and delays the genetic improvement of *A. donax* [13,14,15,16]. However, the fundamental research of *A. donax* genome encountered great difficulties. The ploidy of *A. donax* is still unclear because of the high chromosome number and small chromosome size [17,18]. Similar as another aquatic plant *Phragmites australis* with variable ploidy [19], *A. donax* showed elastic chromosome number. Different chromosome number of *A. donax* were reported by several authors, including 110 [20,21], 108 [22] and 84 chromosomes [23]. The complexity of *A. donax* genome hinders further understanding and the genetic improvement of this valuable energy crop.

The large and complex nature of plant genomes, coupled with varying degrees of ploidy, has long posed challenges for genome assembly in plants. However, advancements in sequencing technologies, such as third-generation long-read sequencing, particularly High Fidelity reads (Hifi) sequencing, along with new assembly algorithms serve as the foundation for the rapid advancements in plant genome sequencing [24]. Hifi sequencing allows for the generation of long-read, high-quality data that can be used to accurately assemble and annotate complex genomes. It has already proven to be a valuable tool in multiple complex plant genomes, such as allohexaploid oat [25], tetraploid potato [26], and autotetraploid alfalfa [27].

Here, we assembled the first two chromosome-scale *A. donax* genome by combining Pacbio SMRT sequencing and high-throughput chromosome conformation capture (Hi-C) technology. These two assemblies will facilitate the better utilization of *A. donax* resourses.

2. Results

2.1. Cytological analysis of A. donax

In the process of collecting domestic *A. donax* germplasm resources, we found that *A. donax* in China can be divided into three main types according to the leaf color, i.e. common, versicolor and emerald green (emerald here after) (Figure 1). Among them, there is little difference in plant architecture, growth period and other aspects between common and versicolor varieties. In contrast, obvious differences can be observed between common and emerald *A. donax* in plant architecture, ear architecture and growth period. Therefore, we selected a common variety 0004 and an emerald variety 0408 for genome assembly.

We first analyzed the karyotype of the two varieties. DAPI staining *in situ* hybridization of telomere repeat sequences showed that a 0004 had 112 (Figure 2A-B) chromosomes, while 0408 had 62 chromosomes (Figure 2D-E), in consistent with the fact that chromosome number in *A. donax* is variable. Moreover, rDNA fluorescence *in situ* hybridization revealed that 6/6 chromosomes in 0004 showed strong hybridization signals of 5SrDNA and 18SrDNA (Figure 2C). In contrast, only 2/2 chromosomes in 0408 showed hybridization signals of 5SrDNA and 18SrDNA (one strong signal and one weak signal) (Figure 2F). These results indicated that 0004 and 0008 may have different chromosome basic number and ploidy.

2.2. High-quality genome assembly of A. donax

Before the genome *de novo* assembly, we conducted k-mer analysis using second generation sequencing (MGI-2000 platform). We first checked for sequence contamination by randomly selecting 100,000 reads and matched against sequences in the nucleotide sequence database. About 79.8% and 84.5% reads of 0004 and 0408 were unmapped in the database, confirming the low contamination of MGI data. Besides, over 2000 read were mapped with *Setaria viridis* and *Digitaria exilis*, indicating the homology of *A. donax* and these two species (Supplementary Figure S1A-B). The estimated genome size of 0004 and 0408 was 1.46 Gb and 3.14 Gb respectively. In addition, we estimated the genome heterozygosity using simulated data of *Arabidopsis*. The genome heterozygosity of 0004 and 0408 is 0.8% and 0.3 % respectively (Supplementary Figure S1C-D).

The genome assembly were performed using long reads generated by PacBio platform. After quality control, we obtained 152.30 Gb reads with N50 16.92 kb for 0004 and 133.18 Gb reads with N50 21.37 kb for 0408. The preliminary assembly of 0408 was 2.97 Gb, in consistent with 3.14 Gb estimated with genome survey. However, assembly of 0004 was 2.24 Gb, much larger than 1.46 Gb. Consider the high heterozygosity of 0004 genome (0.8%), we used Purge_Dups (-f .9) [28] to obtain the 1.41 Gb non-redundant genome. The GC depth analysis showed improvement of genome quality for 0408 after redundancy reduction (Supplementary Figure S2).

We use Benchmarking Universal Single-Copy Orthologs (BUSCO) and CEGMA to

assess the completeness of genome assembly. The percentage of complete BUSCOs was 99.57% and 99.69% in 0004 and 0408 genome (Supplementary Figure S3A-B), and the percentage of complete core gene was 92.74 and 91.53% in 0004 and 0408 genome, confirming the completeness of two genome assembly (Supplementary Figure S3C-D). Besides, to determine the accuracy of the genome assembly, we mapped the MGI sequence to the assemblies. The map rate is 99.60% and 99.96% for 0004 and 0408 respectively, proving the accuracy of the genome assemblies.

2.3. Hi-C scaffolding

Since the chromosome basic number of *A. donax* is still unclear, we first assumed that the two *A. donax* varieties are diploid, thus the chromosome basic number is 56 and 31 for 0004 and 0408 respectively. We performed Hi-C scaffolding and the Hi-C interaction heatmap showed that both 0004 and 0408 are not diploid (Supplementary Figure S4). The number of Hi-C interaction signals for 0004 and 0408 is around 40 and 70 respectively. A hypothesis is that ancestor of all Arundo species is *A. plinii*, whose haploid chromosome number is 36 (n=36) [29]. Therefore, we redid Hi-C scaffolding with n = 36 for 0004 (Figure 3A). Overall, about 99.78% of total sequences were anchored into 36 pseudochromosomes with sizes ranging from 18.58 Mb to 55.91 Mb. The final genome assembly is 1.30 G with Scaffold N50 37.31 Mb (Table 1 and Figure 3C). Based on these results, we speculated normal *A. donax* is triploid (2n = 3x = 108), while 0004 is an aneuploid with four extra chromosomes (3x + 4 =112). The LAI score of assembly constructed by Hi-C scaffolding was 12.63, reaching to the standard of reference quality.

However, for the 0408 genome assembly, the chromosome basic number is obscure. The number of Hi-C interaction signals (about 70) is more than the total chromosome number (62). Therefore, we assumed that 0408 is haploid and redid Hi-C scaffolding with n = 62 for 0408 (Figure 3B). About 97.00% of total sequences were anchored into 62 pseudochromosomes with sizes ranging from 24.80 Mb to 97.92 Mb. The final genome size is 2.86 Gb with scaffold N50 44.86 Mb (Figure 3D). The LAI score of assembly constructed by Hi-C scaffolding was 8.94 for 0408, approaching to the standard of reference quality.

To clarify the relationship between 0004 and 0408, we compared genomes of 0004 and 0408. We identified some chromosome structure variations of the two genomes. More interestingly, the two genomes showed an obvious 1:2 pattern of syntenic depth between 0004 and 0408 (Figure 3E). Therefore, we speculated that 0408 is an allodiploid.

Table 1. Statistics of genome assemblies.

	0004	0408
Assembly features		
Number of scaffolds	65	162
Total size of scaffolds	1299.92 Mb	2770.95 Mb
Longest scaffold	55.91 Mb	97.92 Mb
Shortest scaffold	18.58 Mb	24.80 Mb
Mean scaffold size	20.00 Mb	17.10 Mb
N50 scaffold length	37.31 Mb	44.86 Mb
L50 scaffold count	7	9
Scaffold GC content	44.07%	44.05%
Scaffold N content	0.0002%	0.0003%
Percentage of assembly in scaffolded contigs	99.78%	97.96%
Average number of contigs per scaffold	1.46	1.12
BUSCO (complete)	99.57%	99.69%
LTR Assembly Index (LAI)	12.63	8.94
Gene models		
Number of gene models	74,403	167,586
Mean coding sequence length	1192.62	1,138.43
Mean number of exons per gene	5.31	5.02
Mean exon length	224.79	226.57
Mean intron length	537.64	518.93
Non-protein-coding RNA		
Number of rRNA	2,320	6,305
Number of sRNA	3,118	6,702
Number of regulatory	19	29
Number of tRNA	1,392	3,175

2.4. Repeat elements analysis and gene model prediction

The total length of interspersed repeats is 711.24 Mb (54.71% of the genome) for 0004 and 1627.27 Mb (56.96% of the genome) for 0408. To be specific, a total of 70,102 (0.07% of the genome) and 149,490 (0.07% of the genome) simple repeat sequence (SSR) were identified in 0004 and 0408 genome; a total of 89,378 (0.71% of the genome) and 181,468 (0.78% of the genome) tandem repeat sequences were identified in 0004 and 0408 genome; a total of 1,400,366 (51.54% of the genome) and 2,937,325 (52.91% of the genome) transposable elements (TEs) were identified in 0004 and 0408 genome. The detailed statistics of TEs were listed in Supplementary file (TE).

We performed gene structure prediction by combining transcriptome prediction, homologous protein prediction and ab initio prediction. Firstly, we found that the mapped rate of RNA-seq data to the genome in four tissues were all over 90% (Supplementary table 1), and the mapped rate of Pacbio Isoseq to the genome is 99.86% for 0004 and

99.87% for 0408, further confirming the accuracy of transcriptome data and the genome assemblies. The RNA-seq and Pacbio transcripts were used for gene prediction, resulting in 49,524 / 80,249 predicted genes for 0004 and 0408 respectively. Secondly, we selected five Poaceae plants, including *Saccharum spontaneum*, *Sorghum bicolor*, *Zea mays*, *Triticum aestivum* and *Oryza sativa* for homologous protein prediction. A total of 94,613 / 172,436 genes for 0004 and 0408 were predicted. Thirdly, we performed ab initio prediction, and 78,102 / 174,926 gene models for 0004 and 0408 were predicted. The final gene set were obtained by integrating the above results. In total, 74,403 / 167,586 gene models with average gene length 3,507.41 / 3,226.92 bp, average CDS length 1192.62 / 1,138.43 bp, average exon length 224.79 / 226.57 bp, average intron length 537.64 / 518.93 bp for 0004 and 0408 respectively (Supplementary file (Gene prediction)). Except the gene model, we also predicted non-coding RNA. In total, 2,320 / 6,305 rRNA, 3,118 / 6702 small RNA, 1,392 / 3,175 tRNA were identified in 0004 and 0408 genome. The parameters of the two assemblies were listed in Supplementary file (ncRNA).

2.5. Gene function annotation and evaluation of genome annotation

We predicted the gene function based on five databases, including Non-Reduntant Protein Database (NR), Kyoto Encyclopedia of Gene and Genomes (KEGG), Eukaryotic Orthologous Groups of protein (KOG), GO and Swissprot (Supplementary Figure S5). In total, 67,377 / 152,275 genes were annotated, account for 90.56% / 90.86 % of the 0004 and 0408 genome (Supplementary file (Gene prediction)). The co-annotated gene number is 16,877 / 37,776 for 0004 and 0408 genome (Supplementary Figure S6).

The annotated gene sets were evaluated using BUSCO. Among the 1,614 BUSCO groups, about 98.45% / 97.96% of complete gene elements can be found in the annotated gene set for 0004 and 0408 respectively, indicating that the majority of conservative gene predictions are relatively complete, indicating the high reliability of the gene prediction result. Besides, the proportion of expressed gene in four tissues ranged from 71.09% to 80.34% in 0004 and 67.01% to 73.75% in 0408. Total expressed gene account for 86.60% and 82.53% in 0004 and 0408 (Supplementary file1 (Transcripts)). The gene structure in 0004 and 0408 genome showed similar distribution trend with other Poaceae plants, including gene length, CDS length, exon length, exon number, intron length and intron number (Supplementary Figure S7), demonstrating the reliability of the genome annotations.

3. Materials and Methods

3.1. Plant materials

The two *A. donax* plants were collected from Wuhan, Hubei (0004) and Yueyang, Hunan (0408) respectively. About 3 cm stem with a tiller bud were used as explant to propagate with tissue culture. Seedlings were then acclimatized outside and transplanted in the field.

241 *3.2. Karyotype analysis*

242 Karyotype analysis were conducted followed a previously report [30,31]. Briefly, root tips
243 about 2 cm in length were placed in Nitrous Oxide (0.9-1.0 Mpa) for two hours, and fixed
244 with pre-cold glacial acetic acid. After that, Root tips were digested with cellulase and
245 pectinase at 37°C for 1 hour and crushed using a dissecting needle. Cells suspension were
246 observed under an optical microscope to get the suitable chromosome specimens.
247 Fluorescent *in situ* hybridization probes were prepared using nick translation method.
248 Probes for telomere, 5SrDNA and 18SrDNA are fluorescently-labeled $(TTTAGGG)_6$
249 oligonucleotides, pTa794 and pTa71 plasmids respectively. After hybridization, slides were
250 photographed under a fluorescence microscope.

251 *3.3. Flow cytometry analysis*

252 Karyotype analysis were conducted followed a previously report [32]. Samples were
253 immersed into MG^b dissociation solution (45 mM $MgCl_2 \cdot 6H_2O$, 20 mM MOPS, 30 mM
254 sodium citrate, 1% (W/V) PVP, 0.2% (v/v) Tritonx-100, 10 mM Na 2 EDTA, 20 μL/mL β-
255 mercaptoethanol pH 7.5), chopped with blade and filtered to get the nuclear suspension.
256 The suspension was then incubated in propidium iodide and RNase for 1 h and detected
257 with flow cytometry (BD FACScalibur). Tomato and barley were used as internal reference.

258 *3.4. DNA extraction, Pacbio library preparation and sequencing*

259 High molecular weight genomic DNA was prepared with QIAGEN® Genomic kit (13343)
260 according to the manufacturer's instructions. SMRTbell libraries were constructed PacBio's
261 standard protocol. Briefly, the quality qualified DNA was sheared by g-TUBEs (Covaris,
262 USA), damage repaired and 20-50 kb DNA fragments were screened by the BluePippin
263 (Sage Science, USA) and purified using AMPure PB beads (Agilent technologies, USA).
264 Sequencing was performed on a PacBio Sequel instrument with Sequel II Sequencing Kit
265 2.0. Raw sequencing data were filtered through the SMRTlink 8.0.

266 *3.5. MGISEQ-2000 library preparation and sequencing*

267 Genomic DNA was extracted using CTAB method and subjected to sonication by Covaris.
268 The DNA fragments range from 200-400 bp were selected using Agencourt AMPure XP-
269 Medium kit. After end-repair, 3' adenylated, adapters-ligation and PCR Amplifying, the
270 libraries were prepared using Hieff NGS® Fast-Pace DNA Cyclization Kit (Yeasen,
271 13341ES96) and sequencing on MGISEQ-2000 platform.

272 To facilitate the annotation of the genome, we performed mRNA sequencing on four
273 tissues of the plant including leaf, root, panicle and lateral bud. We purified up to 400 μg of
274 total RNA per sample using the TRIzol-based method (Invitrogen, USA) and subsequently
275 treated it with DNase I. The quality of the extracted RNA was evaluated using a 1% agarose
276 gel and Aglient 2100 bioanalyzer (Agilent Technologies, USA). PolyA mRNA was enriched
277 using the protocol of NEBNext ® Poly(A) mRNA Magnetic Isolation Module (New England

Biolabs #7490 S, USA). RNA sequencing libraries were then prepared using NEBNext ® Ultra TM II RNA Library Prep Kit for Illumina (New England Biolabs #E7775, USA) following the manufacturer's instructions. The RNA library concentration was determined using the Qubit ® 3.0 Fluorometer, and the libraries were then subjected to 150 bp paired-end sequencing on a MGISEQ-2000 platform. In total, we generated 55.8 Gb / 51.9 Gb raw data for the four tissue samples of 0004 and 0408 respectively.

3.6. Estimation of genome size and heterozygosity

The genome size of *A. donax* was estimated based on k-mer analysis using KMC 3 program [33]. The GCE and FindGSE software were used to analyze the 17-mer frequency distribution and genome size estimation [34]. We estimated the genome heterozygosity by combining simulated data with different heterozygosity in *Arabidopsis* genome and the distribution of 17 k-mer data of *A. donax*.

3.7. De novo assembly

The quality-filtered HiFi reads were used for genome assembly with hifiasm (v.0.12) [35] to obtain the preliminary assembly. The completeness of genome was assessed using BUSCO (v.4.0.5) [36] and CEGMA (v.2) [37]. To assess the accuracy of the assembly, all the MGISEQ-2000 reads were mapped to the genome using BWA [38], and SAMtools [39] and bcftools [40] were used to calculate the error rate of genome assembly. Beisides, LTR Assembly Index (LAI) was calculated to assess the assembly quality [41].

3.8. Hi-C scaffolding

To attach the scaffolds to chromosomes, we extracted genomic DNA using young leaves for the Hi-C library, and the sequencing data was obtained via the MGISEQ-2000 platform. To elaborate, we first cut freshly harvested leaves into 2 cm pieces and vacuum infiltrated them with nuclei isolation buffer supplemented with 2% formaldehyde. Crosslinking was stopped by adding glycine and performing additional vacuum infiltration. We then converted the fixed tissue into powder and re-suspended it in nuclei isolation buffer to obtain a nuclei suspension. D*pn*II was used to digest the purified nuclei with 100 units, and after marking with biotin-14-dCTP, biotin-14-dCTP from non-ligated DNA ends was removed through the exonuclease activity of T4 DNA polymerase. The ligated DNA was sheared into 300−600 bp fragments, followed by blunt-end repair and A-tailing before purification through biotin-streptavidin-mediated pull down. Finally, the Hi-C libraries were quantified and sequenced using the MGI-2000 platform.

In total, 375 Gb and 347 Gb paired-end reads were generated for 0004 and 0408. To ensure the quality of the Hi-C raw data, we employed HiC-Pro (v2.8.1), a previously established quality control tool [42]. We first filtered out low-quality sequences (quality scores < 20), adapter sequences, and sequences shorter than 30 bp using fastp [43]. The clean paired-end reads were then mapped to the draft assembled sequence using bowtie2

(v2.3.2) with the parameters "-end-to-end --very-sensitive -L 30" to obtain the unique mapped paired-end reads [44]. We identified and retained valid interaction paired reads using HiC-Pro from the unique mapped paired-end reads for further analysis. Additionally, we filtered out invalid read pairs, including dangling-end, self-cycle, re-ligation, and dumped products, using HiC-Pro. We further clustered, ordered, and oriented the scaffolds onto chromosomes using LACHESIS [45], with parameters CLUSTER_MIN_RE_SITES=100, CLUSTER_MAX_LINK_DENSITY=2.5, CLUSTER NONINFORMATIVE RATIO = 1.4, ORDER MIN N RES IN TRUNK=60, ORDER MIN N RES IN SHREDS=60. Finally, we manually adjusted any placement and orientation errors exhibiting obvious discrete chromatin interaction patterns.

3.9. Repeat Annotation

To identify tandem repeats and transposable elements in the genome, we used GMATA [46] and Tandem Repeats Finder (TRF) software [47]. GMATA identified simple repeat sequences (SSRs), while TRF recognized all tandem repeat elements in the genome. To identify transposable elements, we used a combination of ab initio and homology-based methods. An ab initio repeat library for the genome was predicted using MITE-hunter [48] and RepeatModeler with default parameters, including LTR_FINDER [49], LTRharvest [50], and LTR_retriever [51] for plant genome. The resulting library was aligned to TEclass Repbase (http://www.girinst.org/repbase) to classify each repeat family's type. To identify repeats throughout the genome, we applied RepeatMasker [52] to search for known and novel TEs using the *de novo* repeat library and Repbase TE library. Finally, we collated and combined overlapping transposable elements belonging to the same repeat class.

3.10. Gene Prediction

To predict genes in a repeat-masked genome, we employed three distinct approaches: ab initio prediction, homology search, and reference-guided transcriptome assembly. Homolog prediction was performed using GeMoMa by aligning homologous peptides from related species to the assembly to obtain gene structure information [53]. For RNAseq-based gene prediction, we aligned filtered MGISEQ-2000 reads to the reference genome using STAR [54] (default settings) and then assembled the resulting transcripts with stringtie2 [55]. Open reading frames (ORFs) were predicted using PASA [56]. Ab initio gene prediction was performed with Augustus [57], utilizing the default parameters with the training set. We then employed EVidenceModeler (EVM) [58] to produce an integrated gene set, removing genes with TE using the TransposonPSI package (http://transposonpsi.sourceforge.net/) and filtering miscoded genes. Untranslated regions (UTRs) and alternative splicing regions were identified using PASA based on RNA-seq assemblies. We retained the longest transcripts for each locus, while regions outside of the ORFs were designated as UTRs.

3.11. Functional annotation of gene models

Gene function information, as well as motifs and domains of their proteins, were assigned by comparing them with various publicly available databases, including SwissProt, NR, KEGG, KOG, and Gene Ontology. We utilized the InterProScan program with default parameters to identify the putative domains and GO terms of the genes [59]. We also employed BLASTp to compare the EvidenceModeler-integrated protein sequences against four well-known public protein databases, using an E-value cutoff of 1e-05. We selected the results with the lowest E-value. Results from the searches of these five databases were then concatenated to obtain a comprehensive gene function annotation.

3.12. Annotation of non-coding RNAs (ncRNAs)

For the identification of ncRNA (non-coding RNA), we utilized two main strategies: searching against databases and prediction with models. To predict transfer RNAs (tRNAs), we applied tRNAscan-SE with eukaryote parameters [60]. We employed Infernal cmscan [61] to search the Rfam database for microRNA, rRNA, small nuclear RNA, and small nucleolar RNA detection. Additionally, we utilized RNAmmer to predict the rRNAs and their subunits [62].

3.13. Collinearity analysis

Pairwise alignment between 0004 and 0408 genome were performed using Minimap2 software [63] with parameters: -cx asm10. The paf file were used to obtain the dotplot using R package dotPlotly (https://github.com/tpoorten/dotPlotly/).

4. Discussion

4.1 A. donax is a promising energy crop with complex genome

The rapidly deteriorating environment in earth has made an urgent claim for carbon emission reduction [64]. Using bioenergy to replace part of fossil energy can make remarkable contributions to reducing carbon emissions [65]. However, the land-intensive bioenergy is limited by the finite land resource. Especially in China, because of the huge population size, croplands are inviolable to guarantee the food security. However, except for the cropland, there are about 78 million hectares of marginal lands [66]. Due to climate, terrain, soil and other limiting factors, marginal lands cannot be planted with food crops, while energy crops can often grow on these lands owing to their great environmental adaptabilit. Of all the energy crops, the high yield, extensive adaptability and low inputs make *A. donax* an ideal energy crop.

Here, we generated the first two chromosome-scale assemblies of the *A. donax* genome using HiFi reads and Hi-C technology. According to the LAI score of the assemblies, the assembly of 0004 reached to the criterion of reference quality, while assembly of 0408 also approached to the reference quality (Table 1). Based on the genome, we speculated that

normal *A. donax* is a triploid (2n = 3x = 108), and the aneuploid 0004 were generated because of the abnormal replication or deletion of four chromosomes. In contrast, the genome of emerald *A. donax* is far more complicated. The chromosome base number (about 70) is more than the total chromosome number 62 (Figure 2), and most chromosomes of 0004 can aligned with two chromosomes of 0408. Therefore, we guess emerald *A. donax* is allodiploid (2n = 2x =72), and 0408 is an aneuploid with abnormal chromosome deletion.

4.2 The reference genome will contribute to genetic improvement of A. donax

Clonal propagation crops like potato often showed small genetic gains. Recently, using precise genome design, vigorous F_1 hybrids potatoes has been developed [67]. Similar to potato, *A. donax* is completely agamic and showed a narrow genetic variability. To accelerate genetic improvement of *A. donax* in future, we need to elucidate the mechanism of infertility of *A. donax*. In addition, genetic transformation of *A. donax* encountered difficulties, and no transgenic *A. donax* reported until now. We speculate the difficulty of genetic transformation of *A. donax* lied in the complexity of *A. donax* genome. Several developmental regulators, including BBM, WUS, GRF4 and GIF were reported to improve transformation efficiency in crops like corn and wheat [68]. Therefore, cloning key developmental regulators in *A. donax* may help to break down technical barriers in genetic transformation of *A. donax*. To overcome the above two difficulties, a reference genome is indispensable.

Until now, little is known about the *A. donax* genome, except a few transcriptome studies [69,70,71,72]. Our result showed that *A. donax* has a complex genome that contains a large number of genes, many of which are involved in important biological processes such as growth, development, and stress response. Mining important genes from the reference genome will undoubtedly contribute to genetic improvement of *A. donax*.

In conclusion, *A. donax* is a valuable plant species with numerous characteristics that make it an attractive resource for various applications. Its fast growth, high biomass production, and environmental tolerance make it ideal for bioenergy production and environmental remediation. The plant's genetic diversity also offers many opportunities for genetic improvement and crop breeding. Research on *A. donax* continues to uncover new insights into the plant's biological properties and the potential for its utilization in different industries. Future studies will undoubtedly uncover more valuable information and offer even more exciting opportunities to take advantage of this remarkable plant species.

Authorship contribution statement

Mengmeng Ren: Data analysis and writing. Fupeng Liu and Xiaohong Han: Plant material collection.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

Not applicable.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Reference

1. Corno, L.; Pilu, R.; Adani, F. *Arundo donax* L.: a non-food crop for bioenergy and bio-compound production. *Biotechnol. Adv*. **2014**, *32*, 1535–1549.
2. Jámbor A.; Török, A. The Economics of *Arundo donax*—A Systematic Literature Review. *Sustainability* **2019**, *11*, 4225.
3. Pilu, R.; Manca, A.; Landoni, M. *Arundo donax* as an energy crop: pros and cons of the utilization of this perennial plant. *Maydica* **2013**, *58*.
4. Zhang, J.; Li, Y.; Zhang, C.; Jing, Y. Adsorption of malachite green from aqueous solution onto carbon prepared from *Arundo donax* root. *J. Hazard. Mater*. **2008**, *150*, 774–782.
5. Angelini, L.G., Ceccarini, L.; Bonari, E. Biomass yield and energy balance of giant reed (*Arundo donax* L.) cropped in central Italy as related to different management practices. *Eur. J. Agron*. **2005**, *22*, 375–389.
6. Angelini, L.G.; Ceccarini, L.; Nasso, N.; Bonari, E. Comparison of *Arundo donax* L. and *Miscanthus* x *giganteus* in a long-term field experiment in Central Italy: Analysis of productive characteristics and energy balance. *Biomass & Bioenergy* **2009**, *33*, 635–643.
7. Nasso, N.N.O.D.; Roncucci, N.; Bonari, E. Seasonal Dynamics of Aboveground and Belowground Biomass and Nutrient Accumulation and Remobilization in Giant Reed (*Arundo donax* L.): A Three-Year Study on Marginal Land. *Bioenergy Research* **2013**, *6*, 725–736.
8. Nackley, L.L.; Kim, S.H. A salt on the bioenergy and biological invasions debate: salinity tolerance of the invasive biomass feedstock *Arundo donax*. *GCB Bioenergy* **2015**, *7*, 752–762.
9. Sánchez, E.; Scordia, D.; Lino, G. Salinity and Water Stress Effects on Biomass Production in Different *Arundo donax* L. Clones. *BioEnergy Res*. **2015**, *8*, 1461–1479.
10. Mirza, N.; Mahmood, Q.; Pervez, A.; Ahmad, R.; Farooq, R.; Shah, M.M.; Azim, M.R. Phytoremediation potential of *Arundo donax* in arsenic-contaminated synthetic wastewater. *Bioresource Technol.* **2010**, *101*, 5815–5819.
11. Mirza, N.; Pervez, A.; Mahmood, Q.; Shah, M.M.; Shafqat, M.N. Ecological restoration of arsenic contaminated soil by *Arundo donax* L. *Ecol. Eng.* **2011**, *37*, 1949–1956.
12. Papazoglou, E.G.; Karantounias, G.A.; Vemmos, S.N.; Bouranis D.L. Photosynthesis and growth responses of giant reed (*Arundo donax* L.) to the heavy metals Cd and Ni. *Environ. Int.* **2005**, *31*, 243–249.
13. Ahmad, R.; Liow, P.S.; Spencer. D.F.; Jasieniuk, M. Molecular evidence for a single genetic clone of invasive *Arundo donax* in the United States. *Aquat. Bot.* **2008**, *88*, 113–120.
14. Danelli, T.; Laura, M.; Savona, M.; Landon, M.; Adani, F.; Pilu, R. Genetic Improvement of *Arundo donax* L.: Opportunities and Challenges. *Plants* **2020**, *9*, 1584.
15. Malone, J.M.; Virtue, J.G.; Williams, C.; Preston, C. Genetic diversity of giant reed (*Arundo donax*) in Australia. *Weed Biol. Manag.* **2017**, *17*.
16. Tarin, D.; Pepper, A.E.; Goolsby, J.A.; Moran, P.J.; Arquieta, A.C.; Kirk, A.E.; Manhart, J.R. Microsatellites Uncover Multiple Introductions of Clonal Giant Reed (*Arundo donax*). Invas. *Plant Sci. Mana*. **2013**, *6*, 328–338.

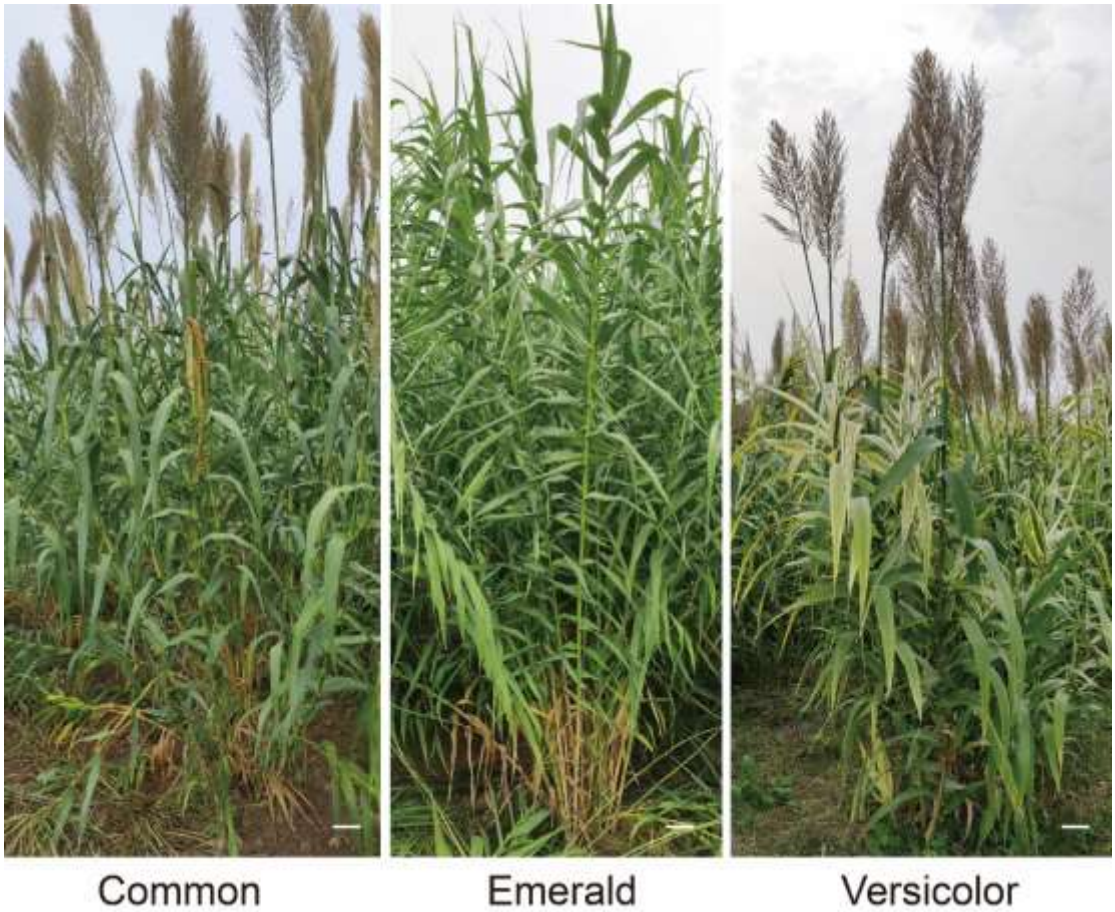
17. Bucci, A.; Cassani, E.; Landoni, M.; Cantaluppi, E.; Pilu, R. Analysis of chromosome number and speculations on the origin of *Arundo donax* L. (Giant Reed). *Cytolo. Genet.* **2013**, *47*, 237–241.
18. Mariani, C.; Cabrini, R.; Danin, A.; Piffanelli, P.; Fricano, A.; Gomarasca, S.; Dicandilo, M.; Grassi, F.; Soave, S. Origin, diffusion and reproduction of the giant reed (*Arundo donax* L.): a promising weedy energy crop. *Ann. Appl. Biol*. **2010**, *157*, 191–202.
19. Clevering, O.A.; Lissner, J. Taxonomy, chromosome numbers, clonal diversity and population dynamics of *Phragmites australis*. *Aquat. Bot*. **1999**, *66*, 249–250.
20. Hunter, A.W.S. A Karyosystematic investigation in Gramineae. *Can. J. Res.* **1934**, *11*, 213–241.
21. Pizzolongo, P. Osservazioni cariologiche su *Arundo donax* e *Arundo plinii*. *Annuali Bot.* **1962**, *27*, 173–187.
22. Christopher, J.; Abraham, A. Studies on the cytology and phylogeny of South Indian grasses I. Subfamilies Bambusoideae, Oryzoideae, Arundinoideae and Festucoideae. *Cytologia* **1971**, *36*, 579–594.
23. Haddadchi, A.; Gross, C.L.; Fatemi, M. The expansion of sterile *Arundo donax* (Poaceae) in southeastern Australia is accompanied by genotypic variation. *Aquat. Bot.* **2013**, *104*, 153–161.
24. Wang, Y.; Yu, J.; Jiang, M.; Lei, W.; Zhang, X.; Tang, H. Sequencing and Assembly of Polyploid Genomes. *Methods Mol. Biol.* **2023**, *2545*, 429–458.
25. Peng, Y., Yan, H., Guo, L., Deng, C.; Wang, C., Wang, Y., Kan, L.; Zhou, P.; Y, K.; Dong, X.; Liu, X.; Su, Z.; Peng, Y.; Zhao, J.; Deng, D.; Xu, Y.; Li, Y.; Jiang, Q.; Li, Y.; Wei, L.; Wang, J.; Ma, J.; Hao, M.; Li, W.; Kang, H.; Peng, Z.; Liu, D.; Jia, J.; Zheng, Y.; Ma, T.; Wei, Y.; Lu, F.; Ren, C. Reference genome assemblies reveal the origin and evolution of allohexaploid oat. *Nat. Genet.* **2022**, *54*, 1248–1258.
26. Sun, H.; Jiao, W.B.; Krause, K.; Campoy, J.A.; Goel, M.; Folz-Donahu, K.; Kukat, C.; Huettel, B.; Schneeberger, K. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat. Genet.* **2022**, *54*, 342–348.
27. Chen, H.; Zeng, Y.; Yang, Y.; Huang, L.; Tang, B.; Zhang, H.; Hao, F.; Liu, W.; Li, Y.; Liu, Y.; Zhang, X.; Zhang, R.; Zhang, Y.; Li, Y.; Wang, K.; He, H.; Wang, Z.; Fan, G.; Yang, H.; Bao, A.; Shang, Z.; Chen, J.; Wang, W.; Qiu Q. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat. Commun.* **2020**, *19*, 2494.
28. Guan, D.; McCarthy, S.A.; Wood, J.; Howe, K.; Wang, Y.; Durbin, R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **2020**, *36*, 2896-2898.
29. Jike, W.; Li, M.; Zadra, L.; Barbaro, E.; Sablok, G.; Bertorelle, G.; Rota-Stabelli, O.; Varotto, C. Phylogenomic proof of Recurrent Demipolyploidization and Evolutionary Stalling of the "Triploid Bridge" in *Arundo* (Poaceae). *Int. J. Mol. Sci.* **2020**, *21*, 5247.
30. Bayani, J.; Squire, J.A. Fluorescence *in situ* Hybridization (FISH). *Current Protocols in Cell Biology* **2004**, Chapter 22.
31. Jiang, J. Fluorescence *in situ* hybridization in plants: recent developments and future applications. *Chromosome Res.* **2019**, *27*, 153–165.
32. Dolezel, J.; Greilhuber, J.; Suda, J. Estimation of nuclear DNA content in plants using

554 flow cytometry. *Nat. Protoc.* **2007**, *2*, 2233–2244.
555 33. Kokot, M.; Dlugosz, M.; Deorowicz, S. KMC 3: counting and manipulating k-mer
556 statistics. *Bioinformatics* **2017**, *33*, 2759-2761.
557 34. Sun, H.; Ding, J.; Piednoël, M.; Schneeberge, K. FindGSE: Estimating genome size
558 variation within human and Arabidopsis using k -mer frequencies. *Bioinformatics* **2018**,
559 *34*: 550–557.
560 35. Cheng, H., Concepcion, G.T.; Feng X.; Zhang, H.; Li, H. Haplotype-resolved *de novo*
561 assembly using phased assembly graphs with hifiasm. *Nat. Methods* **2021**, *18*, 1–6.
562 36. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M.
563 BUSCO: assessing genome assembly and annotation completeness with single-copy
564 orthologs. *Bioinformatics* **2015**, *31*, 3210–3212.
565 37. Parra, G.; Bradnam, K.; Korf, I. CEGMA: a pipeline to accurately annotate core genes
566 in eukaryotic genomes. *Bioinformatics* **2007**, *23*, 1061–1067.
567 38. Li, H.; Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler
568 transform. *Bioinformatics* **2010**, *26*, 589-595.
569 39. Li, H.; Handsaker, B. Wysoker A, et al. The Sequence Alignment/Map format and
570 SAMtools. *Bioinformatics* **2009**, *25*: 2078-2079.
571 40. Danecek, P.; McCarthy, S.A. BCFtools/csq: haplotype-aware variant consequences.
572 *Bioinformatics* **2017**, *33*, 2037–2039.
573 41. Ou, S.; Chen, J.; Ning, J. Assessing genome assembly quality using the LTR Assembly
574 Index (LAI). *Nucleic Acids Res*. **2018**, *46*, e126.
575 42. Servant, N.; Varoquaux, N.; Lajoie, B.R.; Viara, E.; Chen, C.; Vert, J.P.; Heard, E.;
576 Dekker, J.; Barillot, E. HiC-Pro: an optimized and flexible pipeline for Hi-C data
577 processing. *Genome Biol*. **2015**, *16*, 259.
578 43. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
579 *Bioinformatics* **2018**, *34*, i884–i890.
580 44. Langmead, B.; Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*
581 **2012**, *9*, 357-359.
582 45. Burton, J.N.; Adey, A.; Patwardhan, R.P.; Qiu, R.; Kitzman, J.O.; Shendure, J.
583 Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin
584 interactions. *Nat. Biotechnol.* **2013**, *31*, 1119–1125.
585 46. Wang, X.; Wang, L. GMATA: An Integrated Software Package for Genome-Scale SSR
586 Mining, Marker Development and Viewing. *Front. Plant Sci.* **2016**, *7*, 1350.
587 47. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic*
588 *Acids Res.* **1999**, *27*, 573–580.
589 48. Han, Y.; Wessler, S.R. MITE-Hunter: a program for discovering miniature inverted-
590 repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **2010**, *38*:
591 e199.
592 49. Xu, Z.; Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR
593 retrotransposons. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W265–W268.
594 50. Ellinghaus, D.; Kurtz, S.; Willhoeft, U. LTRharvest, an efficient and flexible software
595 for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **2008**, *9*, 18.
596 50. Ou, S.; Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for
597 Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol.* **2018**, *176*,

598 1410–1422.
599 51. Chen, N.S. Using RepeatMasker to identify repetitive elements in genomic sequences.
600 *Curr. Protoc. Bioinformatics* **2004**, *4*, Unit 4.10.
601 52. Keilwagen, J.; Wenk, M.; Erickson, J.L.; Schattat, M.H.; Grau, J.; Hartung, F. Using
602 intron position conservation for homology-based gene prediction. *Nucleic Acids Res.*
603 **2016**, *44*, e89.
604 53. Dobin, A.; Davis, C.A.; Schlesinger, F., Jorg, Drenkow.; Zaleski, C.; Jha, S.; Batut, P.;
605 Chaisson, M.; Gingeras, T.R. STAR: ultrafast universal RNA-seq aligner.
606 *Bioinformatics* **2013**, *29*, 15–21.
607 54. Kovaka, S.; Zimin, A.V.; Pertea, G.M.; Razaghi, R.; Salzberg, S.L.; Pertea, M.
608 Transcriptome assembly from long-read RNA-seq alignments with StringTie2.
609 *Genome Biol.* **2019**, *20*, 278.
610 55. Haas, B.J.; Salzberg, S.L.; Zhu, W.; Pertea, W.; Allen, J.E.; Orvis, J.; White, O.; Buell,
611 C.R.; Wortman, J.R. Automated eukaryotic gene structure annotation using
612 EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.*
613 **2008**, *9*, R7.
614 56. Mario, S.; Burkhard, M. AUGUSTUS: a web server for gene prediction in eukaryotes
615 that allows user-defined constraints. *Nucleic Acids Res.* **2005**, *33* (Web Server issue),
616 W465–W467.
617 57. Haas, B.J.; Salzberg, S.L.; Zhu, W.; Pertea, M.; Allen, J.E; Orvis, J.; White, O.; Buell,
618 C.R.; Wortman, J.R. Automated eukaryotic gene structure annotation using
619 EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome. Biol.*
620 **2008**, *R7*.
621 58. Zdobnov, E.M.; Apweiler, R. InterProScan--an integration platform for the signature-
622 recognition methods in InterPro. *Bioinformatics* **2001**, *17*, 847–848.
623 59. Chan, P.P.; Lin, B.Y.; Mak, A.J. tRNAscan-SE: a program for improved detection of
624 transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997, 25, 955–964.
625 60. Nawrocki, E.P.; Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches.
626 *Bioinformatics* **2013**, *29*, 2933-2935.
627 61. Lagesen, K.; Hallin, P.; Rødland, E.A.; Stærfeldt, H.H.; Rognes, T.; Ussery, D.W.
628 RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids*
629 *Res*. **2007**, *35*, 3100-3108.
630 62. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**,
631 *34*, 3094–3100.
632 63. Patz, J.A.; Frumkin, H.; Holloway, T.; Vimont, D.J.; Haines, A. Climate
633 Change:challenges and opportunities for global health. *JAMA* **2014**, *312*, 1565-1580.
634 64. Walter, V.R.; Mariam, K.A.; Christopher B.F. The future of bioenergy. *Glob Chang Biol.*
635 **2020,** *26*, 274–286.
636 65. Tang, Y.; Xie, J.S.; Geng, S. Marginal Land-based Biomass Energy Production in
637 China. *J. Integr. Plant Biol.* **2010**, *52*, 112–21.
638 66. Zhang, C.; Yang, Z.; Tang, D.; Zhu, Y.; Wang, P.; Li, D.; Zhu, G.; Xiong, X.; Shang, Y.;
639 Li, C.; Huang, S. Genome design of hybrid potato. Cell 2021, 184, 3873–3883.e12.
640 67. Chen, Z.; Debernardi, J.M.; Dubcovsky, J.; Gallavotti, A. Recent advances in crop
641 transformation technologies. *Nat. Plants* **2022**, *8*, 1343–1351

68. Evangelistella, C.; Valentini, A.; Ludovisi, R.; Firrincieli, A.; Fabbrini, F.; Scalabrin, S.; Cattonaro, F.; Morgante, M.; Mugnozza, G.S.; Keurentjes, J.J.B.; Harfouche, A. *De novo* assembly, functional annotation, and analysis of the giant reed (*Arundo donax* L.) leaf transcriptome provide tools for the development of a biofuel feedstock. Biotechnol. *Biofuels* **2017**, *10*, 138.
69. Fu, Y.; Poli, M.; Sablok, G.; Wang, B.; Liang, Y.; Porta, N.L.; Velikova, V.; Loreto, F., Li, M., Varotto, C. Dissection of early transcriptional responses to water stress in *Arundo donax* L. by unigene-based RNA-seq. *Biotechnol. Biofuels* **2016**, *9*, 54.
70. Sablok, G.; Fu, Y.; Bobbio, V.; Laura, M.; Rotino, G.L.; Bagnaresi, P.; Allavena, A.; Velikova, V.; Viola, R.; Loreto, F.; Li, M.; Varotto, C. Fuelling genetic and metabolic exploration of C3 bioenergy crops through the first reference transcriptome of *Arundo donax* L.. *Plant Biotechnol. J*. **2014**, *12*, 554–567.
71. Sicilia, A.; Testa, G.; Santoro, D.F.; Cosentino, S.L.; Piero, A.R.L. RNASeq analysis of giant cane reveals the leaf transcriptome dynamics under long-term salt stress. *BMC Plant Biol*. **2019**, *19*.

683 **Figure legend**



684
685 **Figure 1.** Morphology of the common, emerald and versicolor *A. donax*. Pictures were
686 taken in September 2022 in Shishou, Hubei province (E112.40°, N29.73°). Giant panicles
687 can be observed in common and versicolor *A. donax*, while no panicle was observed in
688 emerald *A. donax*. Bar = 50 cm.

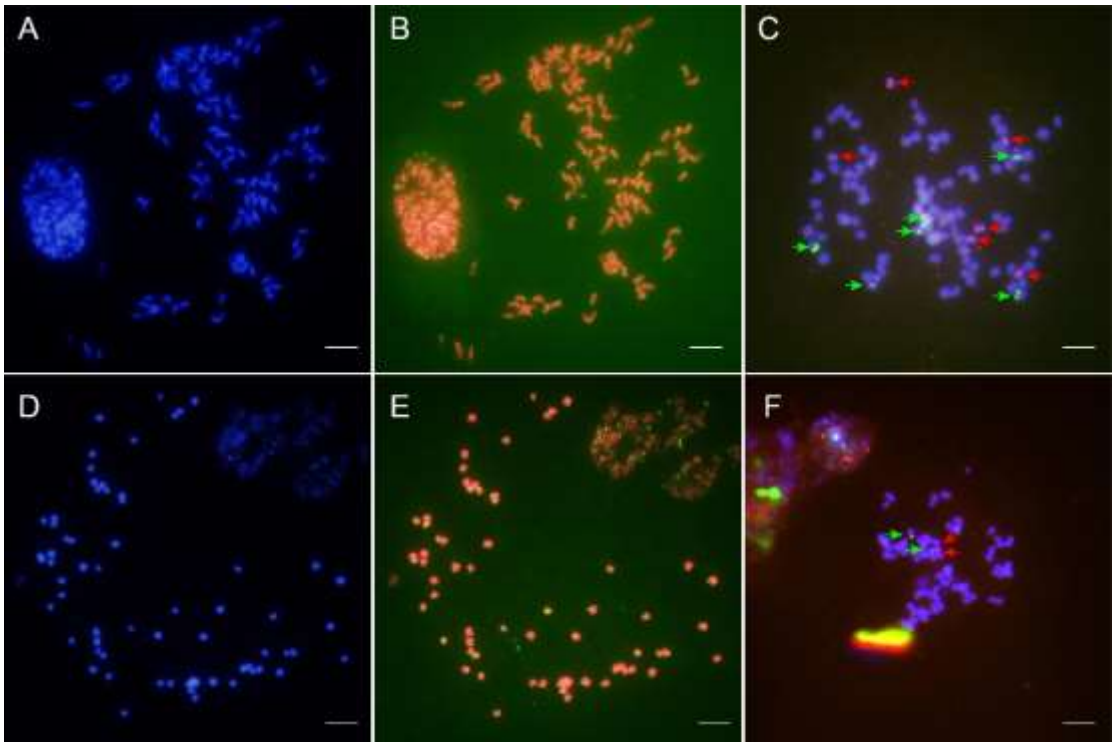
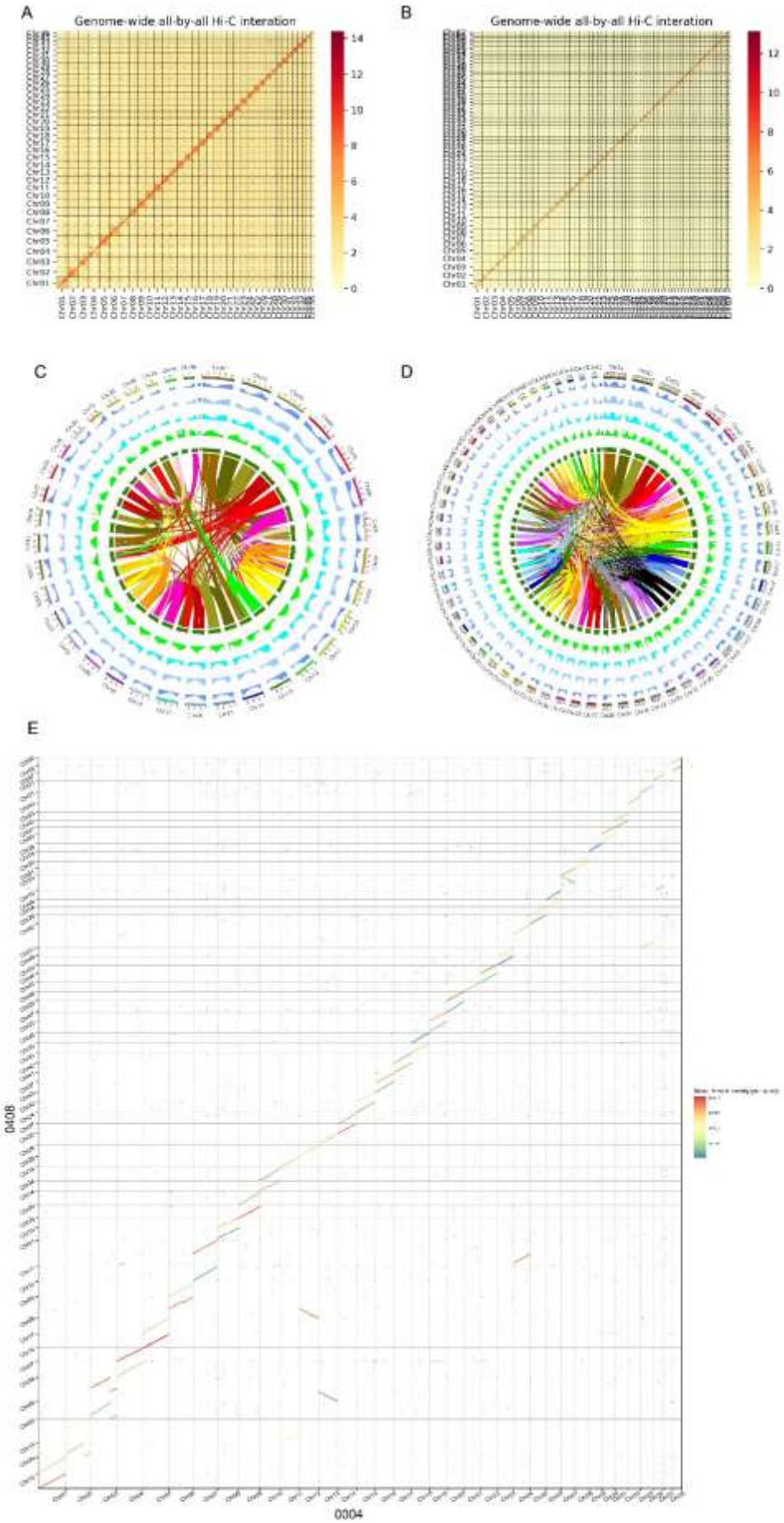
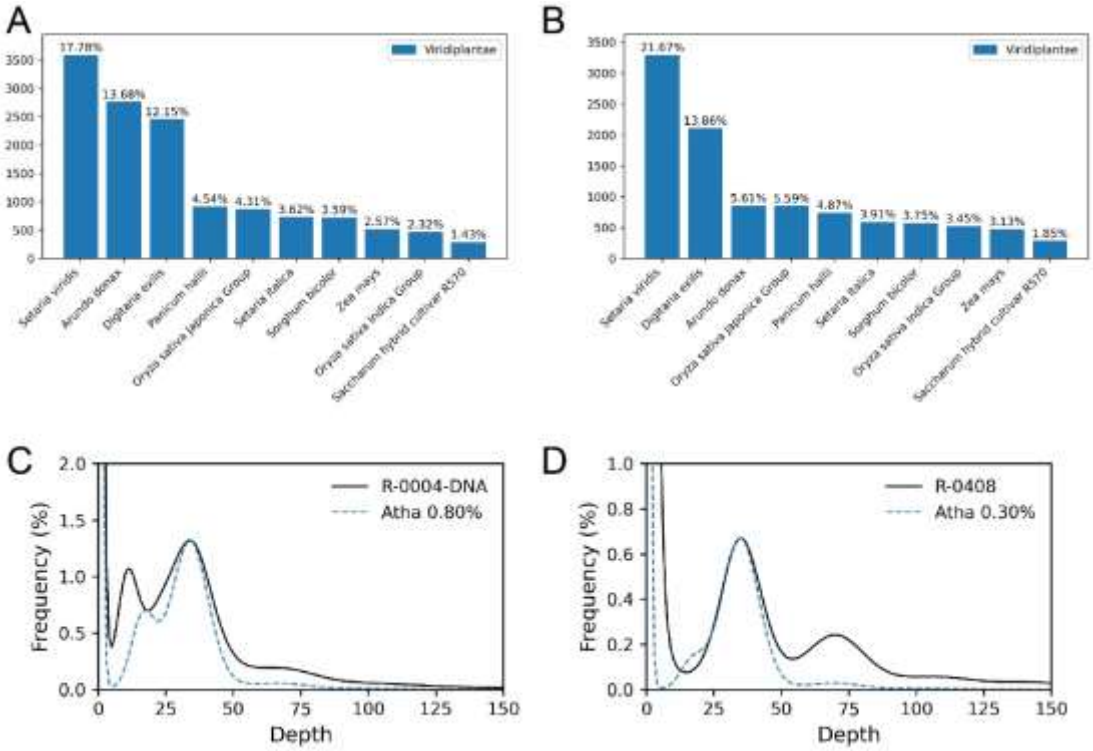


Figure 2. Karyotype analysis of 0004 and 0408. (A, D) DAPI staining of chromosomes of 0004 (A) and 0408 (D). (B, E) *In situ* hybridization of telomere repeat sequences of 0004 (B) and 0408 (E). (C, F) rDNA fluorescence *in situ* hybridization of chromosomes of 0004 (C) and 0408 (F). Red and green arrows denote the 5SrDNA and 18SrDNA hybridization signals. Bar = 50 μm.


A
Genome-wide all-by-all Hi-C interation
B
Genome-wide all-by-all Hi-C interation
C
D
E
0408
0004

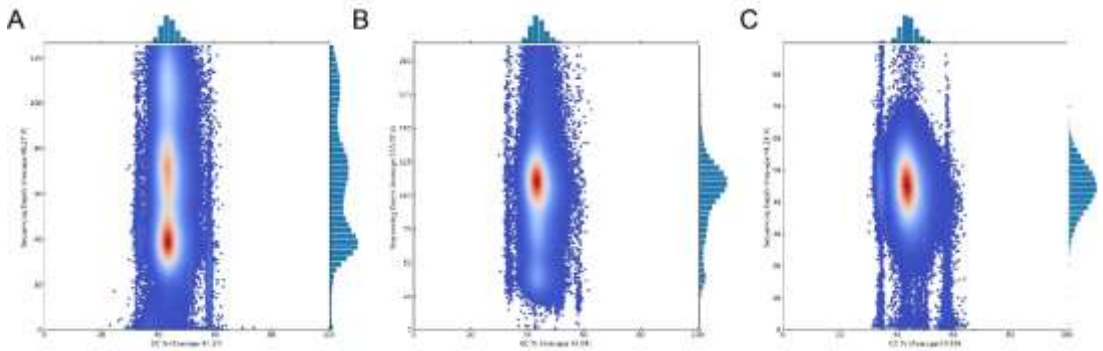
695

696 **Figure 3.** Overview of the *A. donax* genome assemblies. (A, B) Genome-wide Hi-C map
697 of 0004 (A) and 0408 (B). (C, D) Circos plots of 0004 (C) and 0408 (D) genome assemblies.
698 The density of GC content (a), repeated sequence distribution (b), gene density (c),
699 functional genes (d), expressed genes (e) were calculated using 100kb non-overlap
700 window. The innermost lines represent inter-chromosomal synteny. (E) Macrosynteny
701 dotplot of 0004 and 0408 chromosomes.



702

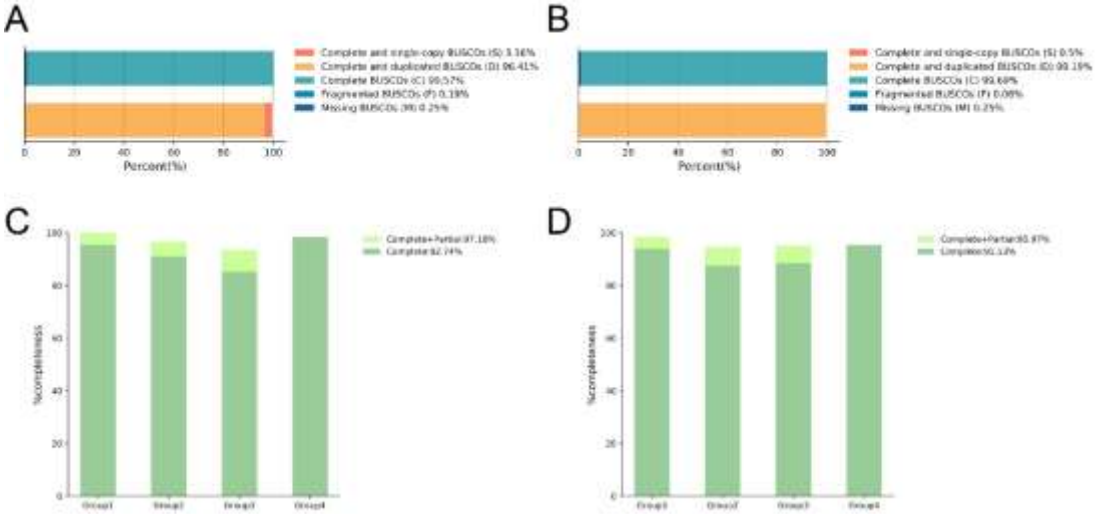
703 **Supplementary Figure S1.** Genome survey of 0004 and 0408. (A-B) Main species
704 mapped in nucleotide sequence database using MGI sequence of 0004 (A) and 0408 (B).
705 (C-D) K-mer distribution curve and heterozygosity simulation curve of 0004 (C) and 0408
706 (D). Horizontal axis is the K-mer depth, vertical axis is frequency of the K-mer depth.
707 Genome of *Arabidopsis thaliana* (Atha) was used to estimate the genome heterozygosity
708 of *A.donax*.



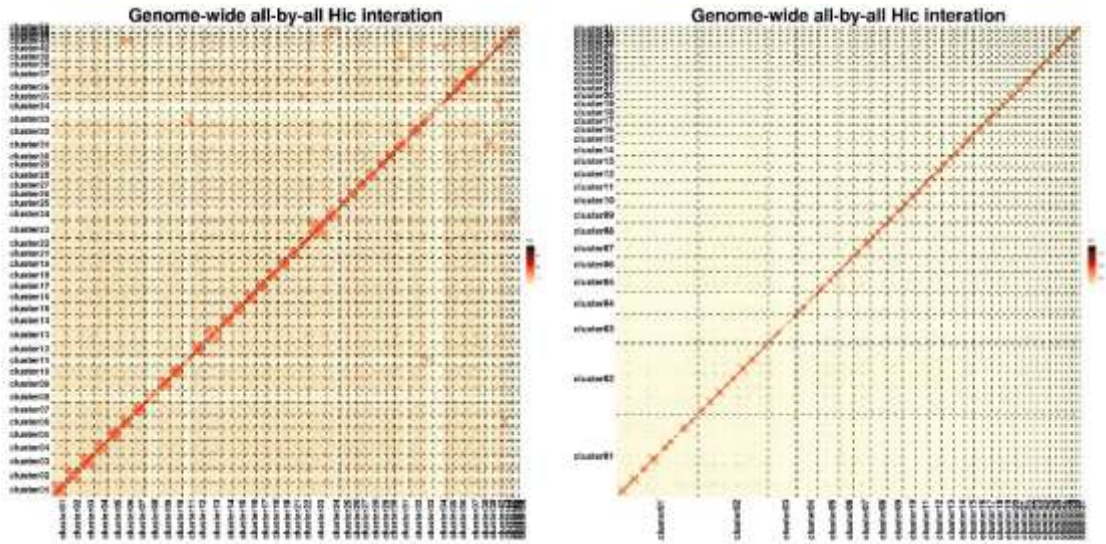
709

710 **Supplementary Figure S2.** GC depth distribution of 0004 genome before elimination of
711 redundancy (A), 0004 genome after elimination of redundancy and 0408 genome. The
712 horizontal axis represents the GC content, and the vertical axis represents the Depth. Data

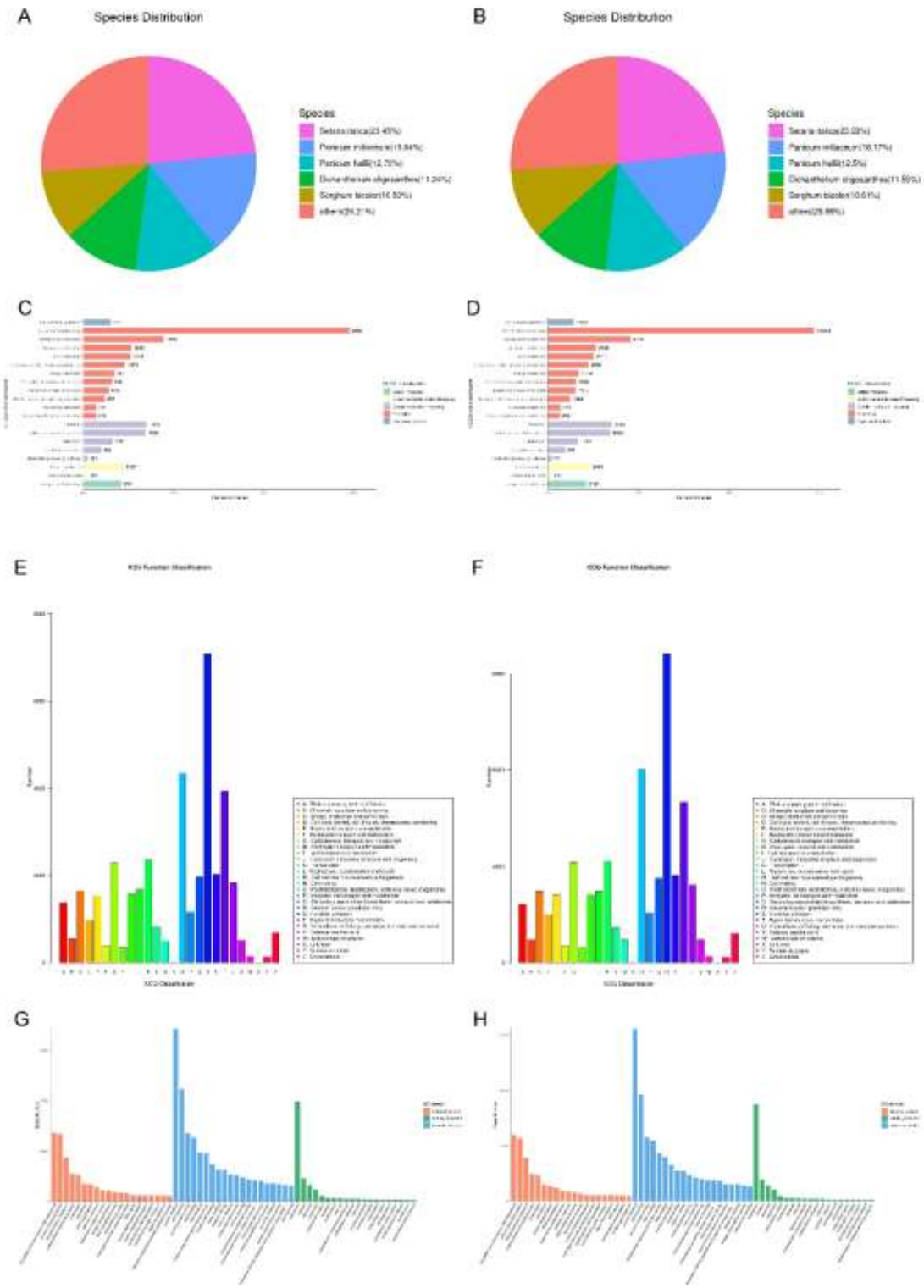
were obtained in 10 Kb successively window.



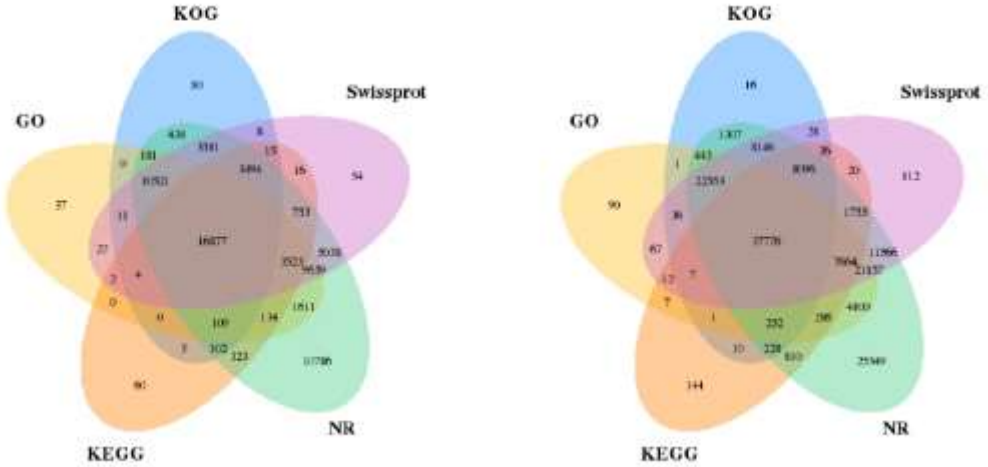
Supplementary Figure S3. Assessment of genome assembly completeness. (A-B) Genome assembly completeness was assessed using BUSCO for 0004 (A) and 0408 (B). (C-D) Conserved gene was assessed using CEGMA for 0004 (C) and 0408 (D).



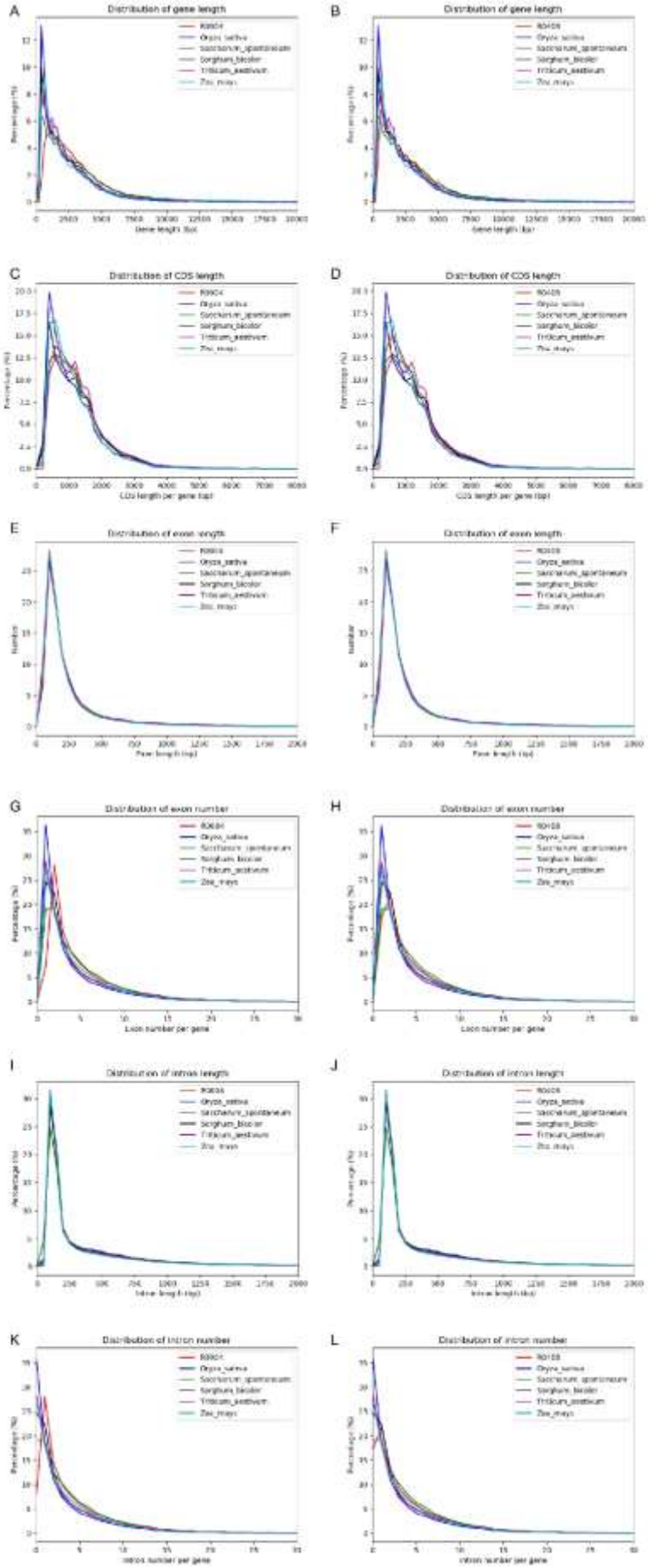
Supplementary Figure S4. Heatmap of primary Hi-C scaffolding of 0004 (left) and 0408 (right). The chromosome basic number was assumed to 56 for 0004 and 31 for 0408. The heat map coordinates represent chromosomes, and the color of each point represents the log value of the corresponding genome bin pair interaction intensity, which increases sequentially from yellow to black.



Supplementary Figure S5. Genome annotation based on public databases. (A-B) Species distribution for NR-based annotation for 0004 (A) and 0408 genome (B). (C-D) KEGG classification of 0004 (C) and 0408 genome (D) annotation. (E-F) KOG function classification of 0004 (E) and 0408 (F) genome annotation. (G-H) GO domain of 0004 (G) and 0408 (H) genome annotation.



Supplementary Figure S6. Statistics of genome annotation for 0004 (left) and 0408 genome based on five databases, including Non-Reduntant Protein Database (NR), Kyoto Encyclopedia of Gene and Genomes (KEGG), Eukaryotic Orthologous Groups of protein (KOG), GO and Swissprot.


A
Distribution of gene length
B
Distribution of gene length
C
Distribution of CDS length
D
Distribution of CDS length
E
Distribution of exon length
F
Distribution of exon length
G
Distribution of exon number
H
Distribution of exon number
I
Distribution of intron length
J
Distribution of intron length
K
Distribution of intron number
L
Distribution of intron number

736 **Supplementary Figure S7.** Comparative statistics of gene information of related species.
737 Comparison of gene information of 0004 (left column) and 0408 (right column) with five
738 Poaceae plants (*Oryza sativa*, *Saccharum spontaneum*, *Sorghum bicolor*, *Zea mays* and
739 *Triticum aestivum*) including distribution of gene length (A-B), CDS length (C-D), exon
740 length (E-F), exon number (G-H), intron length (I-J) and intron number (K-L).
741
742 **Supplementary table 1** Statistics of RNA-seq data

Sample name	Total reads	Total mapped	Unique mapped	Multiple mapped
0004-root	75,570,094	69,694,640 (92.23%)	67,979,118 (89.96%)	1,715,522 (2.27%)
0004-leaf	73,310,846	68,078,324 (92.86%)	66,040,232 (90.08%)	2,038,092 (2.78%)
0004-panicle	121,085,174	109,619,650 (90.53%)	107,277,684 (88.60%)	2,341,966 (1.93%)
0004-bud	102,532,136	95,570,548 (93.21%)	93,230,426 (90.93%)	2,340,122 (2.28%)
0408-root	112,487,910	110,342,744 (98.09%)	95,485,032 (84.88%)	14,857,712 (13.21%)
0408-leaf	81,606,088	79,939,116 (97.96%)	68,384,754 (83.80%)	11,554,362 (14.16%)
0408-panicle	62,558,910	59,462,320 (95.06%)	50,570,542 (80.84%)	8,891,778 (14.22%)
0408-bud	89,793,602	87,915,064 (97.91%)	75,257,540 (83.81%)	12,657,524 (14.10%)

743