Chromosome-scale Genome Assembly of the Alloenneaploid Arundo donax

Mengmeng Ren
Wuhan Rundo Biotechnology CO., Ltd
Xiaohong Han
Wuhan Rundo Biotechnology Co., Ltd
Fupeng Liu
Wuhan Rundo Biotechnology Co., Ltd
Daohong Wu
Wuhan Rundo Biotechnology Co., Ltd
Hai Peng (penghai138@163.com)
Jianghan University https://orcid.org/0000-0001-7827-3322

Research Article

Keywords: A. donax, genome assembly, allopolyploid, comparative genomics

Posted Date: January 12th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3831980/v1

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 the first chromosome-scale assembly of *A. donax* genome using Pacbio SMRT sequencing and Hi-C technology. The genome size of this assembly is 1.30 Gb with contig N50 33.15 Mb. A total of 74,403 gene models were predicted, of which over 90% of genes were functionally annotated. Karyotype analysis and synteny analysis revealed that *A. donax*is an alloenneaploid (3n = 9x = 108). Comparative genome analysis indicated that *A. donax* has undergone strong gene family expansion and two whole-genome duplication events during evolution. Based on the genome assembly, we mined numerous salinity stress related genes using public RNA-seq data. The genome assembly we provided in this study will enhance genetic understanding and promote the genetic improvement of *A. donax*.

Key message

The chromosome-scale genome assembly provides a reference genome and uncover the allopolyploid feature of *Arundo donax.*

Introduction

Arundo donax L. (*A. donax*) is a tall perennial grass 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 (Corno et al. 2014; Jámbor & Török. 2019; Pilu et al. 2013; Zhang et al. 2008). One of the significant characteristics of *A. donax* is its high productivity and fast growth rate. The plant matures rapidly and can produce large quantities of biomass in a short time, making it an attractive crop for bioenergy production (Angelini et al. 2005; Angelini et al. 200 2009; Nasso et al. 2013). Additionally, *A. donax* is highly tolerant of a broad range of environmental conditions, including drought, excess water, and high salinity (Nackley & Kim, 2015; Sánchez et al. 2015). 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 (Mirza et al. 2010; Mirza et al. 2011). *A. donax* is effective in removing heavy metals and other pollutants from contaminated sites, making it a valuable resource in environmental restoration (Papazoglou et al. 2005). 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 limits the genetic variability and delays the genetic improvement of *A. donax* (Ahmad et al. 2008; Danelli et al. 2020; Malone et al. 2017; Tarin et al. 2013). 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 size of each chromosome (Bucci et al. 2013; Mariani et al. 2010). Similar to another aquatic plant *Phragmites australis* (Clevering & Lissner, 1999), *A. donax* showed elastic chromosome number with variable ploidy. Different chromosome number of *A. donax* were reported by several authors, including 110 (Hunter, 1934; Pizzolongo, 1962), 108 (Christopher & Abraham, 1971), and 84 chromosomes (Haddadchi et al. 2013). 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. Fortunately, advancements in sequencing technologies, such as third-generation long-read sequencing, particularly High Fidelity (HiFi) reads sequencing, along with new assembly algorithms serve as the foundation for the rapid advancements in plant genome sequencing (Wang et al. 2023). 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 decoding complex plant genomes, such as allohexaploid oat (Peng et al. 2022), tetraploid potato (Sun et al. 2022), and autotetraploid alfalfa (Chen et al. 2020).

Here, we assembled the first chromosome-scale *A. donax* genome by combining Pacbio SMRT sequencing and high-throughput chromosome conformation capture (Hi-C) technology. This assembly will facilitate the better understanding of *A. donax* genomics and better utilization of *A. donax* resources.

Materials and methods

Plant materials

The *A. donax* plants were collected from Wuhan, Hubei, China (114.28 E, 30.79 N). 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.

Karyotype analysis

Karyotype analysis was conducted followed a previously report (Bayani & Squire, 2004; Jiang et al. 2019). Briefly, root tips about 2 cm in length were placed in Nitrous Oxide (0.9-1.0 Mpa) for two hours, and fixed with pre-cold glacial acetic acid. After that, Root tips were digested with cellulase and pectinase at 37°C for 1 h and crushed using a dissecting needle. Cells suspension was observed under an optical microscope to get the suitable chromosome specimens. Fluorescent *in situ* hybridization probes were prepared using the nick translation method. Probes for telomere, 5SrDNA and 18SrDNA are fluorescently-labeled $(TTTAGGG)_6$ oligonucleotides, pTa794 and pTa71 plasmids respectively. After hybridization, slides were photographed under a fluorescence microscope.

DNA extraction, Pacbio library preparation and sequencing

High molecular weight genomic DNA was prepared with QIAGEN® Genomic kit (13343) according to the manufacturer's instructions. SMRTbell libraries constructed PacBio's standard protocol. Briefly, the quality

qualified DNA was sheared by g-TUBEs (Covaris, USA), damage repaired and 20–50 kb DNA fragments were screened by the BluePippin (Sage Science, USA) and purified using AMPure PB beads (Agilent technologies, USA). Sequencing was performed on a PacBio Sequel instrument with Sequel II Sequencing Kit 2.0. Raw sequencing data were filtered through SMRTlink 8.0.

MGISEQ-2000 library preparation and sequencing

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

To facilitate the annotation of the genome, we performed mRNA sequencing on four plant tissues including leaf, root, panicle and lateral bud. We purified up to 400 μg of total RNA per sample using the TRIzol-based method (Invitrogen, USA), and subsequently treated it with DNase I. PolyA mRNA was enriched using the protocol of NEBNext ® Poly(A) mRNA Magnetic Isolation Module (New England Biolabs #7490 S, USA). RNA sequencing libraries were then prepared using Hieff NGS® Ultima Dual-mode mRNA Library Prep Kit (Yeasen, 12309ES96) 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 raw data for the four tissue samples respectively.

Estimation of genome size and heterozygosity

The genome size of *A. donax* was estimated based on k-mer analysis using KMC3 program (Kokot et al. 2017). The GCE and FindGSE software were used to analyze the 17-mer frequency distribution and perform genome size estimation (Sun et al. 2018). We estimated the genome heterozygosity by combining simulated data with different heterozygosity in the *Arabidopsis* genome and the distribution of 17 k-mer data of *A. donax*.

De novo assembly

The quality-filtered HiFi reads were used for genome assembly with hifiasm (v.0.12) (Cheng et al. 2021) to obtain the preliminary assembly. The completeness of genome was assessed using BUSCO (v.4.0.5) (Simão et al. 2015) and CEGMA (v.2) (Parra et al. 2007). To assess the accuracy of the assembly, all the MGISEQ-2000 reads were mapped back to the genome using BWA (Li & Durbin, 2010), and SAMtools (Li et al. 2009) and bcftools (Danecek & McCarthy, 2017) were used to calculate the error rate of genome assembly. Besides, LTR Assembly Index (LAI) was calculated to assess the assembly quality (Ou et al. 2018).

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. *Dpn*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 MGISEQ-2000 platform.

In total, 375 Gb paired-end reads were generated for Hi-C. To ensure the quality of the Hi-C raw data, we employed HiC-Pro (v2.8.1), a previously established quality control tool (Servant et al. 2015). We first filtered out low-quality sequences (quality scores < 20), adapter sequences, and sequences shorter than 30 bp using fastp (Chen et al. 2018). 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 (Langmead & Salzberg, 2012). 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 (Burton et al. 2013), 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.

Repeat element Annotation

To identify tandem repeats and transposable elements in the genome, we used GMATA (Wang & Wang, 2016) and Tandem Repeats Finder (TRF) software (Benson, 1999). 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 (Han & Wessler, 2010) and RepeatModeler with default parameters, including LTR_FINDER (Xu & Wang, 2007), LTRharvest (Ellinghaus et l., 2008), and LTR_retriever (Ou & Jiang, 2018) 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 (Chen, 2004) 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.

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 (Keilwagen et al. 2016). For RNAseq-based gene prediction, we aligned filtered MGISEQ-2000 reads to the reference genome using STAR (Dobin et al. 2013) (default settings) and then assembled the resulting transcripts with stringtie2 (Kovaka et al. 2019). Open reading frames (ORFs) were predicted using PASA (Haas et al. 2008). Ab initio gene prediction was performed with Augustus (Mario et al. 2005), utilizing the default parameters with the training set. We then employed EVidenceModeler (EVM) (Haas et al. 2008) 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 assembly. We retained the longest transcripts for each locus, while regions outside the ORFs were designated UTRs.

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 (Zdobnov & Apweiler, 2001). 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.

Annotation of non-coding RNAs (ncRNAs)

For the identification of ncRNA, we utilized two main strategies: searching against databases and prediction with models. To predict transfer RNAs (tRNAs), we applied tRNAscan-SE was applied with eukaryote parameters (Chan et al. 1997), andInfernal cmscan (Nawrocki & Eddy, 2007) was employed 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 (Lagesen et al. 2017).

Comparative genomic analysis

Collinearity analysis was carried out using MCScan (Python version) software. Dotplots for genome pairwise synteny was visualized with the command ‘python -m jcvi.graphics.dotplot’.

To investigate the evolutionary history of *A. donax*, six grass families, *Oryza sativa*, *Zea mays*, *Sorghum bicolor*, *Brachypodium distachyon*, *Setaria italica*, *Saccharum spontaneum* and one dicot plant *Arabidopsis thaliana* were used for orthologous analysis, phylogenetic analysis and gene family expansion and contraction analysis using OrthoVenn3 (Sun et al. 2023) with default parameters.

Whole-genome duplication analysis was performed using wgdi (Sun et al. 2022). The synonymous substitution values (Ks) of syntenic gene pairs were calculated with default parameters and visualized using R program.

RNA-seq analysis

The raw RNA seq sequencing data (GSE121552, GSE125104) was downloaded from NCBI. We used fastp to filter the low-quality sequences, adapter sequences, and short sequences. The filtered reads were maped to the reference genome using hisat2 (Kim et al. 2019) to obtain sam file, followed by using samtools to obtain the sorted bam file. The expression matrix was calculated using featureCounts (Liao et al. 2014). DeSeq2 (Love et al. 2014) was used to identify differentially expressed genes (FDR < 0.05 and $\log_2$(fold change) > 2). Gene enrichment analysis was performed using TBtools-II (Chen et al. 2023) and visualized using R program.

Results

Karyotype and k-mer analysis of *A. donax*

Chromosome section and DAPI staining *in situ* hybridization of telomere repeat sequences showed that this plant had 108 chromosomes (Fig. 1A-B), in consistent with previous report (Christopher & Abraham, 1971). The rDNA fluorescence *in situ* hybridization revealed that 6/6 chromosomes showed strong hybridization signals of 5S rDNA and 18S rDNA (Fig. 1C). To clarify the ploidy feature and genome size of *A. donax*, we conducted k-mer analysis using next generation sequencing reads (MGI-2000 platform). The 17-mer frequency distribution curve showed two obvious peaks on 11.4 and 34.2 depth. The triple relationship of the two peaks hinted that *A. donax* is a triploid, consistent with a precious assertion (Jike et al. 2022). Besides, the estimated genome size of *A. donax* was 1.46 Gb, and genome heterozygosity was 0.8%.

High-quality genome assembly of *A. donax*

The long reads used for genome assembly were generated by PacBio platform. After quality control, we obtained 152.30 Gb reads with N50 16.92 kb. The preliminary assembly of *A. donax* was 2.24 Gb, much larger than 1.46 Gb. Considering the relatively high heterozygosity of 0004 genome (0.8%), we used Purge_Dups (-f .9) (Guan et al. 2020) to obtain the 1.41 Gb non-redundant genome, occupying 96.6% of the estimated genome (1.46 Gb). The GC depth analysis showed an improvement of genome quality for *A. donax* after redundancy reduction (Supplementary Fig. S1).

We used Benchmarking Universal Single-Copy Orthologs (BUSCO) and CEGMA to assess the completeness of genome assembly. The percentage of complete BUSCOs was 99.57% in the assembly (Supplementary Fig. S2A), and the percentage of complete core genes was 92.74% (Supplementary Fig. S2B), confirming the completeness of the genome assembly. Besides, we mapped the MGI short reads back to the assembly, the alignment rate is 99.60%, proving the accuracy of the genome assembly.

As we mentioned above, *A. donax* is predicted to be a triploid, thus the haploid chromosome number is 36. Therefore, we performed Hi-C scaffolding with n = 36, and about 99.78% of total sequences were anchored into 36 pseudo chromosomes with sizes ranging from 18.58 Mb to 55.91 Mb (Fig. 2A). The

final genome assembly is 1.30 Gb with scaffold N50 37.31 Mb (Table 1 and Fig. 2B). The LTR Assembly Index (LAI) score was 12.63, reaching to the standard of reference quality. Overall, the genome assembly is a haploid genome with high quality.

Table 1
Statistics of genome assemblies.

	Statistics
Assembly features	
Number of scaffolds	65
Total size of scaffolds	1299.92 Mb
Longest scaffold	55.91 Mb
Shortest scaffold	18.58 Mb
Mean scaffold size	20.00 Mb
N50 scaffold length	37.31 Mb
L50 scaffold count	7
Scaffold GC content	44.07%
Scaffold N content	0.0002%
Percentage of assembly in scaffolded contigs	99.78%
Average number of contigs per scaffold	1.46
BUSCO (complete)	99.57%
LTR Assembly Index (LAI)	12.63
Gene models	
Number of gene models	74,403
Mean coding sequence length	1192.62
Mean number of exons per gene	5.31
Mean exon length	224.79
Mean intron length	537.64
Non-protein-coding RNA	
Number of rRNA	2,320
Number of sRNA	3,118
Number of regulatory	19
Number of tRNA	1,392

Repeat elements analysis and gene model prediction

We first analyzed the interspersed repeats in *A. donax* genomes. The total length of interspersed repeats is 711.24 Mb (54.71% of the genome). To be specific, a total of 70,102 (0.07% of the genome) simple repeat sequence (SSR), 89,378 (0.71% of the genome) tandem repeat sequences, 1,400,366 (51.54% of the genome) transposable elements (TEs) were identified in the genome. The detailed statistics of TEs were listed in Supplementary file2 (TE).

We performed gene structure prediction by combining transcriptome prediction, homologous protein prediction, and ab initio prediction. Firstly, we found that the alignment rate of RNA-seq data to the genome in four tissues were all over 90% (Supplementary file 3), and the alignment rate of Pacbio Isoseq to the genome is 99.86%, further confirming the accuracy of transcriptome data and the genome assembly. The RNA-seq and Pacbio transcripts were used for gene prediction, resulting in 49,524 predicted genes. 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 genes were predicted. Thirdly, we performed ab initio prediction, and 78,102 gene models were predicted. The final gene set was obtained by integrating the above results. In total, 74,403 gene models with average gene length 3,507.41 bp, average CDS length 1192.62 bp, average exon length 224.79 bp, average intron length 537.64 bp (Supplementary file2 (Gene prediction)). Except the gene model, we also predicted non-coding RNA. In total, 2,320 rRNA, 3,118 small RNA, 1,392 tRNA were identified in the genome. The parameters of the assembly were listed in Supplementary file2 (ncRNA).

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 Fig. S3). In total, 67,377 genes were annotated, accounting for 90.56% of the genomes (Supplementary file2 (Gene prediction)). The co-annotated gene number is 16,877 for the genomes (Supplementary Fig. S4).

The annotated gene sets were evaluated using BUSCO. Among the 1,614 BUSCO groups, about 98.45% of complete gene elements can be found in the annotated gene set, indicating that the majority of conservative gene predictions are relatively complete and confirming the high reliability of the gene prediction result. Besides, the proportion of expressed genes in four tissues ranged from 71.09–80.34%. Total expressed genes account for 86.60% of the whole gene sets (Supplementary file2 (Transcripts)). The gene structure in *A. donax* genome showed similar distribution trend with other Poaceae plants, including gene length, CDS length, exon length, exon number, intron length and intron number (Supplementary Fig. S5), demonstrating the reliability of the genome annotations.

A. donax is an alloenneaploid

Based on the protein sequence of the genome assembly, we performed intra-genomic comparison within *A. donax*. The discontinuous synteny chromosome segments revealed that *A. donax* undergone multiple

chromosome rearrangement during evolutionary process. Interestingly, most single chromosome segment can be aligned to two other chromosome segments (Fig. 3A). Besides, synteny analysis of *A. donax* and *S. italica* showed a 1:3 syntenic relationship (Fig. 3B and Supplementary Fig. S6). Based on these results, we speculated that *A. donax* is an enneaploid.

To determine whether *A. donax* is autoenneaploid or alloenneaploid, we used SubPhaser to split the subgenomes of *A. donax.* The k-mer based heatmap showed that the chromosomes were clustered to two groups, in which subgenome A has 12 chromosomes and subgenome B has 24 chromosomes (Fig. 3C). Therefore, our results jointly proved that *A. donax* is alloenneaploid, and the karyotype is AAABBBBBB (3n = 9x = 108). The whole-genome duplication (WGD) analysis showed that *A. donax* undergone two WGD event, one is the ancient ρ event shared by Poaceae plants (Wang et al. 2015), another is a recent burst WGD event occurred ~ 13.5 MYA (Fig. 3D).

Gene family clustering analysis of *A. donax*

To investigate the genome evolutionary history of *A. donax*, gene family clustering was carried out using *A. donax*, six other Gramineae species (*Oryza sativa*, *Zea mays*, *Sorghum bicolor*, *Brachypodium distachyon*, *Setaria italica* and *Saccharum spontaneum*) and a dicotyledon *Arabidopsis thaliana*. A total of 21,162 gene clusters were identified in *A. donax* genomes, in which 9,063 clusters were shared by all the above species, and 1,989 clusters were unique to *A. donax* (Fig. 4A). A total of 121 Single-copy orthologs shared by *Arundo* and six other grass plants were used for phylogenetic analysis and divergence time estimation, which showed that subfamily *Arundo*, *Setaria and Panicum* shared a common ancestor ~ 49.5 million years ago (MYA) (Fig. 4B).

A. donax undergone dramatic gene family expansion during evolution (Fig. 4C). The 611 expanded gene families were enriched in GO terms like “response to water deprivation”, “response to oxidative stress”, “response to osmotic stress”, “response to cold” and “response to heat” (Fig. 4D), hinting that *A. donax* emerged from the grass family because of the server environment in earth at that time.

Salt stress response gene mining of *A. donax*

Two previous studies had identified multiple salt stress response genes using RNA-seq (Angelo et al. 2019, 2020), while the analysis was based on transcript assembly and offer limited information of specific genes. To deeply mine salt stress response genes, we reanalyzed the public RNA-seq data using the genome assembly above. The mapping rate of RNA-seq data was 67.0%~71.1% and 71.1%~74.4% (Supplementary file 3), furthering proving the accuracy of the assembly. Gene expression heatmap using transcripts per million values (TPM) showed that one of the two studies showed low data consistency among the biological duplications (Supplementary Fig. S7), while another study showed better data quality (Fig. 5A). Therefore, we used the data with high consistency which contained two gradients of salt treatment (server and extreme) to perform the following analysis.

A total of 3471 differential expression genes (DEGs) were identified in three comparisons. In details, 956 DEGs were identified in CK versus severe (240 up-regulated and 746 down-regulated), 2875 DEGs were identified in CK versus extreme (1119 up-regulated and 1756 down-regulated), 1395 DEGs were identified in severe versus extreme (607 up-regulated and 787 down-regulated) (Fig. 5B and supplementary file 4). Next, 584 DEGs (overlap of CK_severe and CK_extreme) were used for GO enrichment analysis. Interestingly, top 15 enrichment pathways contained “response to water deprivation”, “response to water”, “response to salt”, proving that these DEGs were indeed response to salt stress (Fig. 5C).

Discussion

A. donax is a promising energy crop with complex genome

The rapidly deteriorating environment on earth has made an urgent claim for carbon emission reduction (Patz et al. 2014). Using bioenergy to replace part of fossil energy can make remarkable contributions to reducing carbon emissions (Walter et al. 2020). However, land-intensive bioenergy is limited by the finite land resource. Especially in China, because of the huge population size, croplands are inviolable to guarantee food security. However, there are about 78 million hectares of marginal lands except for the cropland (Tang et al. 2010). 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 adaptability. Of all the energy crops, the high yield, extensive adaptability and low inputs make *A. donax* an ideal energy crop.

The chromosome number and ploidy of *A. donax i*s controversial. Here, we generated the first chromosome-scale assembly of the *A. donax* genome using HiFi reads and Hi-C technology. Combine karyotype analysis, K-mer analysis and Hi-C scaffolding, we inferred that *A. donax* is an alloenneaploid (3n = 9x = 108), which was supported by a recent study (Jike et a al., 2020). A recent burst WGD event occurred ~ 13.5 MYA in *A. donax* (Fig. 3D), while it is puzzling that a how a single WGD event can form such complicated chromosome components. To unveil the mystery of chromosome evolutionary process *A. donax*, more *Arundo* genus or Arundinoideae plants need to be studied.

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

Clonal propagation crops like potatoes often showed small genetic gains. Using precise genome design, vigorous F_1 hybrids potatoes has been developed recently (Zhang et al. 2021). Similar to potato, *A. donax* is completely agamic and shows a narrow genetic variability. To accelerate the genetic improvement of *A. donax* in the future, we need to elucidate the mechanism of infertility of *A. donax*. In addition, the genetic transformation of *A. donax* encounters difficulties, and no transgenic *A. donax* has been truly reported until now. Several developmental regulators, including BBM, WUS, GRF4 and GIF were reported to improve transformation efficiency in crops like corn and wheat (Chen et al. 2022). 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 for a few transcriptome studies (Evangelistella et al. 2017; Fu et al. 2016; Sablok et al. 2014; Sicilia et al. 2019). We reanalyzed one of the RNA-seq public data and mined multiple salt stress response genes. Among the DEGs, 25 genes were annotated to be response to salt, water and water deprivation (Fig. 5C and supplementary file 4), which showed obvious connection with salt stress related pathways and are potential targets for genetic improvement. This result proved that analyzing RNA-seq data with the genome assembly here provided a feasible approach to mine specific functional gene, and will undoubtedly contribute to the genetic improvement of *A. donax*. Future research on *A. donax* will continues to uncover new insights into the plant's biological properties and the potential for its utilization in different industries. It will undoubtedly uncover more valuable information and offer even more exciting opportunities to take advantage of this remarkable plant species.

Declarations

Author Contributions: Daohong Wu and Hai Peng: Project design, writing review and editing. Mengmeng Ren: Data analysis and writing. Fupeng Liu and Xiaohong Han: Plant material collection, experiment and data collection. All authors have read and agreed to the published version of the manuscript.

Funding: First-class discipline construction funding (Jianghan University, 2023XKZ024); New varieties and patent protection project of forestry plant (National Forestry and Grassland Administration, KJZXXP202314)

Data Availability Statement: The data presented in this study are available on request from the corresponding author. Raw sequence data will be available online later.

Conflicts of Interest: The authors declare no conflict of interest.

ORCID

Mengmeng Ren: https://orcid.org/my-orcid?orcid=0000-0001-8124-6915

Hai Peng: https://orcid.org/my-orcid?orcid=0000-0001-7827-3322

References

1. Ahmad R, Liow PS, Spencer DF, Jasieniuk M (2008) Molecular evidence for a single genetic clone of invasive Arundo donax in the United States. Aquatic Bot 88:113–120. https://doi.org/10.1016/j.aquabot.2007.08.015

2. Angelini LG, Ceccarini L, Bonari E (2005) Biomass yield and energy balance of giant reed (*Arundo donax* L.) cropped in central Italy as related to different management practices. Eur J Agron, 22:375–389. https://doi.org/10.1016/j.eja.2004.05.004
3. Angelini LG, Ceccarini L, Nasso N, Bonari E (2009) 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 Bioenerg 33:635–643. https://doi.org/10.1016/j.biombioe.2008.10.005
4. Bayani J, Squire JA (2004) Fluorescence *in situ* Hybridization (FISH). Current Protocols in Cell Biology Chapter 22. https://doi.org/10.1002/0471143030.cb2204s23
5. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580. https://doi.org/10.1093/nar/27.2.573
6. Bucci A, Cassani E, Landoni M, Cantaluppi E, Pilu R (2013) Analysis of chromosome number and speculations on the origin of *Arundo donax* L. (Giant Reed). Cytol Genet 47:237–241. https://doi.org/10.3103/S0095452713040038
7. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J (2013) Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nat Biotechnol 31:1119–1125. https://doi.org/10.1038/nbt.2727
8. Chan PP, Lin BY, Mak AJ (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964. https://doi.org/10.1093/nar/25.5.955
9. Chen C, Wu Y, Li J, Wang X, Zeng Z, Xu J, Liu Y, Feng J, Chen H, He Y, Xia R (2023) TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. Mol Plant 16:1733–1742. https://doi.org/ 10.1016/j.molp.2023.09.010
10. 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 (2020) Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nat Commun 19:2494. https://doi.org/10.1038/s41467-020-16338-x
11. Chen NS (2004) Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics 4:Unit 4.10. https://doi.org/10.1002/0471250953.bi0410s05
12. Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560
13. Chen Z, Debernardi JM, Dubcovsky J, Gallavotti A (2022) Recent advances in crop transformation technologies. Nat Plants 8:1343–1351 https://doi.org/10.1038/s41477-022-01295-8
14. Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021) Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. Nat Methods 18:1–6. https://doi.org/10.1038/s41592-020-01056-5
15. Christopher J, Abraham A (1971) Studies on the cytology and phylogeny of South Indian grasses I. Subfamilies Bambusoideae, Oryzoideae, Arundinoideae and Festucoideae. Cytologia 36:579–594. https://doi.org/10.1508/cytologia.36.579

16. Clevering OA, Lissner J (1999) Taxonomy, chromosome numbers, clonal diversity and population dynamics of Phragmites australis. Aquat Bot 66:249–250. https://doi.org/10.1016/S0304-3770(00)00094-2
17. Corno L, Pilu R, Adani F (2014) *Arundo donax* L.: a non-food crop for bioenergy and bio-compound production. Biotechnology Advances, 32, 1535–1549. https://doi.org/10.1016/j.biotechadv.2014.10.006
18. Danecek P, McCarthy SA (2017) BCFtools/csq: haplotype-aware variant consequences. Bioinformatics 33:2037–2039. https://doi.org/10.1093/bioinformatics/btx100
19. Danelli T, Laura M, Savona M, Landon M, Adani F, Pilu R (2020) Genetic Improvement of *Arundo donax* L.: Opportunities and Challenges. Plants 9:1584. https://doi.org/10.3390/plants9111584
20. Dobin A, Davis CA, Schlesinger F, Jorg D, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. https://doi.org/10.1093/bioinformatics/bts635
21. Dolezel J, Greilhuber J, Suda J (2007) Estimation of nuclear DNA content in plants using flow cytometry. Nat Protocols 2:2233–2244. https://doi.org/10.1038/nprot.2007.310
22. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics 9:18. https://doi.org/10.1186/1471-2105-9-18
23. Evangelistella C, Valentini A, Ludovisi R, Firrincieli A, Fabbrini F, Scalabrin S, Cattonaro F, Morgante M, Mugnozza GS, Keurentjes JJB, Harfouche A (2017) *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 Biofuel 10:138. https://doi.org/10.1186/s13068-017-0828-7
24. Fu Y, Poli M, Sablok G, Wang B, Liang Y, Porta NL, Velikova V, Loreto F, Li M, Varotto C (2016) Dissection of early transcriptional responses to water stress in *Arundo donax* L. by unigene-based RNA-seq. Biotechnology Biofuels 9:54. https://doi.org/10.1186/s13068-016-0471-8
25. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R (2020) Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics 36:2896–2898. https://doi.org/10.1101/729962
26. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9:R7. https://doi.org/10.1186/gb-2008-9-1-r7
27. Haddadchi A, Gross CL, Fatemi M (2013) The expansion of sterile *Arundo donax* (Poaceae) in southeastern Australia is accompanied by genotypic variation. Aquat Bot 104:53–161. https://doi.org/10.1016/j.aquabot.2012.07.006
28. Han Y, Wessler SR (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38:e199. https://doi.org/
29. Hunter AWS (1934) A Karyosystematic investigation in Gramineae. Canadian Journal of Research 11:213–241. https://doi.org/10.1139/cjr34-087

30. Jámbor A, Török A (2019) The Economics of *Arundo donax*—A Systematic Literature Review. Sustainability 11:4225. https://doi.org/10.3390/su11154225
31. Jia K, Wang Z, Wang L, Li G, Zhang W, Wang X, Xu F, Jiao S, Zhou S, Liu H, Ma Y, Bi G, Zhao W, El-Kassaby YA, Porth I, Li G, Zhang R, Mao J (2022) SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific k-mers. New Phytol 235:801–809. https://doi.org/10.1111/nph.18173
32. Jiang J (2019) Fluorescence *in situ* hybridization in plants: recent developments and future applications. Chromosome Res 27:153–165. https://doi.org/10.1007/s10577-019-09607-z
33. Jike W, Li M, Zadra N, Barbaro N, Sablok G, Bertorelle G, Rota-Stabelli O, Varotto C (2020) Phylogenomic proof of Recurrent Demipolyploidization and Evolutionary Stalling of the "Triploid Bridge" in *Arundo* (Poaceae). Int J of Mol Sci 21:5247. https://doi.org/10.3390/ijms21155247
34. Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F (2016) Using intron position conservation for homology-based gene prediction. Nucleic Acids Res 44:e89. https://doi.org/10.1093/nar/gkw092
35. Kim D, Paggi JM, Park J, Bennett C, Salzberg SL (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907–915. https://doi: 10.1038/s41587-019-0201-4
36. Kokot M, Dlugosz M, Deorowicz S (2017) KMC 3: counting and manipulating k-mer statistics. Bioinformatics 33:2759–2761. https://doi.org/10.1093/bioinformatics/btx30
37. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M (2019) Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol 20:278. https://doi.org/10.1186/s13059-019-1910-1
38. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35: 3100–3108. https://doi.org/10.1093/nar/gkm160
39. Langmead B, Salzberg S (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923
40. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589-595. https://doi.org/10.1093/bioinformatics/btp698
41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352
42. Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930. https://doi: 10.1093/bioinformatics/btt656
43. Love MI, Huber M, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi: 10.1186/s13059-014-0550-8.

44. Malone JM, Virtue JG, Williams C, Preston C (2017) Genetic diversity of giant reed (*Arundo donax*) in Australia. Weed Biol and Manag 17. https://doi.org/10.1111/wbm.12111
45. Mariani C, Cabrini R, Danin A, Piffanelli P, Fricano A, Gomarasca S, Dicandilo M, Grassi F, Soave S (2010) Origin, diffusion and reproduction of the giant reed (*Arundo donax* L.): a promising weedy energy crop. Ann Appl Biol 157:191–202. https://doi.org/10.1111/j.1744-7348.2010.00419.x
46. Mario S, Burkhard M (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 33 (Web Server issue): W465–W467. https://doi.org/10.1093/nar/gki458
47. Mirza N, Mahmood Q, Pervez A, Ahmad R, Farooq R, Shah MM, Azim MR (2010) Phytoremediation potential of *Arundo donax* in arsenic-contaminated synthetic wastewater. Bioresource Technol 101:5815–5819. https://doi.org/10.1016/j.biortech.2010.03.012
48. Mirza N, Pervez A, Mahmood Q, Shah MM, Shafqat MN (2011) Ecological restoration of arsenic contaminated soil by *Arundo donax* L. Ecol Eng 37:1949–1956. https://doi.org/10.1016/j.ecoleng.2011.07.006
49. Nackley LL, Kim SH (2015) A salt on the bioenergy and biological invasions debate: salinity tolerance of the invasive biomass feedstock *Arundo donax*. GCB Bioenergy 7:752–762. https://doi.org/10.1111/gcbb.12184
50. Nasso NNOD, Roncucci N, Bonari E (2013) 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. Bioenerg Res 6:725–736. https://doi.org/10.1007/s12155-012-9289-9
51. Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29:2933–2935. https://doi.org/10.1093/bioinformatics/btt509
52. Ou S, Chen J, Ning J (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res 46:e126. https://doi.org/10.1093/nar/gky730
53. Ou S, Jiang N (2018) LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol 176:1410–1422. https://doi.org/10.1104/pp.17.01310
54. Papazoglou EG, Karantounias G A, Vemmos SN, Bouranis DL (2005) Photosynthesis and growth responses of giant reed (*Arundo donax* L.) to the heavy metals Cd and Ni. Environ Int 31:243–249. https://doi.org/10.1016/j.envint.2004.09.022
55. Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–1067. https://doi.org/10.1093/bioinformatics/btm071
56. Patz JA, Frumkin, H, Holloway T, Vimont DJ, Haines A (2014) Climate Change:challenges and opportunities for global health. JAMA 312:1565–1580. https://doi.org/10.1001/jama.2014.13186
57. 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 (2022) Reference genome assemblies reveal the origin and

evolution of allohexaploid oat. Nat Genet 54:1248–1258. https://doi.org/10.1038/s41588-022-01127-7

58. Pilu R, Manca A, Landoni M (2013) *Arundo donax* as an energy crop: pros and cons of the utilization of this perennial plant. Maydica 58.
59. Pizzolongo P (1962) Osservazioni cariologiche su *Arundo donax* e *Arundo plinii*. Annuali Bot 27:173–187.
60. Sablok G, Fu Y, Bobbio V, Laura M, Rotino GL, Bagnaresi P, Allavena A, Velikova V, Viola R, Loreto F, Li M, Varotto C (2014) Fuelling genetic and metabolic exploration of C3 bioenergy crops through the first reference transcriptome of *Arundo donax* L. Plant Biotechnol J 12:554–567. https://doi.org/10.1111/pbi.12159
61. Sánchez E, Scordia D, Lino G (2015) Salinity and Water Stress Effects on Biomass Production in Different *Arundo donax* L. Clones. Bioenergy Res 8:1461–1479. https://doi.org/10.1007/s12155-015-9652-8
62. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C, Vert JP, Heard E, Dekker J, Barillot E (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol 16:259. https://doi.org/10.1186/s13059-015-0831-x
63. Sicilia A, Testa G, Santoro DF, Cosentino SL, Piero ARL (2019) RNASeq analysis of giant cane reveals the leaf transcriptome dynamics under long-term salt stress. BMC Plant Biol 19. https://doi.org/10.1186/s12870-019-1964-y
64. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351
65. Sun J, Lu F, Luo Y, Bie L, Xu L, Wang Y (2023) OrthoVenn3: an integrated platform for exploring and visualizing orthologous data across genomes. Nucleic Acids Res 51(W1):W397–W403. https://doi.org/10.1093/nar/gkad313
66. Sun P, Jiao B, Yang Y, Shan L, Li T, Li X, Xi Z, Wang X, Liu J (2022) WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. Mol Plant 15:1841–1851. https://doi.org/10.1016/j.molp.2022.10.018
67. Sun H, Ding J, Piednoël M, Schneeberge K (2018) FindGSE: Estimating genome size variation within human and Arabidopsis using k-mer frequencies. Bioinformatics 34:550–557. https://doi.org/10.1093/bioinformatics/btx637
68. Sun H, Jiao WB, Krause K, Campoy JA, Goel M, Folz-Donahu K, Kukat C, Huettel B, Schneeberger K (2022) Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. Nat Genet 54:342–348. https://doi.org/10.1038/s41588-022-01015-0
69. Tang Y, Xie JS, Geng S (2010) Marginal Land-based Biomass Energy Production in China. J Int Plant Biol 52:112–121. https://doi.org/10.1111/j.1744-7909.2010.00903.x
70. Tarin D, Pepper AE, Goolsby JA, Moran PJ, Arquieta AC, Kirk AE, Manhart JR (2013) Microsatellites Uncover Multiple Introductions of Clonal Giant Reed (*Arundo donax*). Invas Plant Sci Mana 6:328–

338. https://doi.org/10.1614/ipsm-d-12-00085.1

71. Walter VR, Mariam KA, Christopher BF (2020) The future of bioenergy. Global Change Biol 26:274–286. https://doi.org/10.1111/gcb.14883
72. Wang X, Wang L (2016) GMATA: An Integrated Software Package for Genome-Scale SSR Mining, Marker Development and Viewing. Front Plant Sci 7:1350. https://doi.org/10.3389/fpls.2016.01350
73. Wang X, Wang J, Jin D, Guo H, Lee T, Liu T, Paterson AH (2015) Genome Alignment Spanning Major Poaceae Lineages Reveals Heterogeneous Evolutionary Rates and Alters Inferred Dates for Key Evolutionary Events. Mol Plant 8:885–898. https://doi: 10.1016/j.molp.2015.04.004
74. Wang Y, Yu J, Jiang M, Lei W, Zhang X, Tang H (2023) Sequencing and Assembly of Polyploid Genomes. Methods in Molecular Biology 2545:429–458. https://doi.org/10.1007/978-1-0716-2561-3_23
75. Xu Z, Wang H (2010) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35 (Web Server issue):W265–W268. https://doi.org/10.1093/nar/gkm286
76. Zdobnov EM, Apweiler R (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17:847–848. https://doi.org/10.1093/bioinformatics/17.9.847
77. Zhang C, Yang Z, Tang D, Zhu Y, Wang P, Li D, Zhu G, Xiong X, Shang Y, Li C, Huang S (2021) Genome design of hybrid potato. Cell 184:3873–3883.e12. https://doi.org/10.1016/j.cell.2021.06.006
78. Zhang J, Li Y, Zhang C, Jing Y (2008) Adsorption of malachite green from aqueous solution onto carbon prepared from *Arundo donax* root. J Haz Mat 150:774–782. https://doi.org/10.1016/j.jhazmat.2007.05.036

div id="Sec17" class="Section2">

Karyotype and k-mer analysis of *A. donax*

Chromosome section and DAPI staining *in situ* hybridization of telomere repeat sequences showed that this plant had 108 chromosomes (Fig. 1A-B), in consistent with previous report (Christopher & Abraham, 1971). The rDNA fluorescence *in situ* hybridization revealed that 6/6 chromosomes showed strong hybridization signals of 5S rDNA and 18S rDNA (Fig. 1C). To clarify the ploidy feature and genome size of *A. donax*, we conducted k-mer analysis using next generation sequencing reads (MGI-2000 platform). The 17-mer frequency distribution curve showed two obvious peaks on 11.4 and 34.2 depth. The triple relationship of the two peaks hinted that *A. donax* is a triploid, consistent with a precious assertion (Jike et al. 2022). Besides, the estimated genome size of *A. donax* was 1.46 Gb, and genome heterozygosity was 0.8%.

High-quality genome assembly of *A. donax*

The long reads used for genome assembly were generated by PacBio platform. After quality control, we obtained 152.30 Gb reads with N50 16.92 kb. The preliminary assembly of *A. donax* was 2.24 Gb, much larger than 1.46 Gb. Considering the relatively high heterozygosity of 0004 genome (0.8%), we used Purge_Dups (-f .9) (Guan et al. 2020) to obtain the 1.41 Gb non-redundant genome, occupying 96.6% of the estimated genome (1.46 Gb). The GC depth analysis showed an improvement of genome quality for *A. donax* after redundancy reduction (Supplementary Fig. S1).

We used Benchmarking Universal Single-Copy Orthologs (BUSCO) and CEGMA to assess the completeness of genome assembly. The percentage of complete BUSCOs was 99.57% in the assembly (Supplementary Fig. S2A), and the percentage of complete core genes was 92.74% (Supplementary Fig. S2B), confirming the completeness of the genome assembly. Besides, we mapped the MGI short reads back to the assembly, the alignment rate is 99.60%, proving the accuracy of the genome assembly.

As we mentioned above, *A. donax* is predicted to be a triploid, thus the haploid chromosome number is 36. Therefore, we performed Hi-C scaffolding with n = 36, and about 99.78% of total sequences were anchored into 36 pseudo chromosomes with sizes ranging from 18.58 Mb to 55.91 Mb (Fig. 2A). The final genome assembly is 1.30 Gb with scaffold N50 37.31 Mb (Table 1 and Fig. 2B). The LTR Assembly Index (LAI) score was 12.63, reaching to the standard of reference quality. Overall, the genome assembly is a haploid genome with high quality.

Table 1
Statistics of genome assemblies.

	Statistics
Assembly features	
Number of scaffolds	65
Total size of scaffolds	1299.92 Mb
Longest scaffold	55.91 Mb
Shortest scaffold	18.58 Mb
Mean scaffold size	20.00 Mb
N50 scaffold length	37.31 Mb
L50 scaffold count	7
Scaffold GC content	44.07%
Scaffold N content	0.0002%
Percentage of assembly in scaffolded contigs	99.78%
Average number of contigs per scaffold	1.46
BUSCO (complete)	99.57%
LTR Assembly Index (LAI)	12.63
Gene models	
Number of gene models	74,403
Mean coding sequence length	1192.62
Mean number of exons per gene	5.31
Mean exon length	224.79
Mean intron length	537.64
Non-protein-coding RNA	
Number of rRNA	2,320
Number of sRNA	3,118
Number of regulatory	19
Number of tRNA	1,392

Repeat elements analysis and gene model prediction

We first analyzed the interspersed repeats in *A. donax* genomes. The total length of interspersed repeats is 711.24 Mb (54.71% of the genome). To be specific, a total of 70,102 (0.07% of the genome) simple repeat sequence (SSR), 89,378 (0.71% of the genome) tandem repeat sequences, 1,400,366 (51.54% of the genome) transposable elements (TEs) were identified in the genome. The detailed statistics of TEs were listed in Supplementary file2 (TE).

We performed gene structure prediction by combining transcriptome prediction, homologous protein prediction, and ab initio prediction. Firstly, we found that the alignment rate of RNA-seq data to the genome in four tissues were all over 90% (Supplementary file 3), and the alignment rate of Pacbio Isoseq to the genome is 99.86%, further confirming the accuracy of transcriptome data and the genome assembly. The RNA-seq and Pacbio transcripts were used for gene prediction, resulting in 49,524 predicted genes. 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 genes were predicted. Thirdly, we performed ab initio prediction, and 78,102 gene models were predicted. The final gene set was obtained by integrating the above results. In total, 74,403 gene models with average gene length 3,507.41 bp, average CDS length 1192.62 bp, average exon length 224.79 bp, average intron length 537.64 bp (Supplementary file2 (Gene prediction)). Except the gene model, we also predicted non-coding RNA. In total, 2,320 rRNA, 3,118 small RNA, 1,392 tRNA were identified in the genome. The parameters of the assembly were listed in Supplementary file2 (ncRNA).

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 Fig. S3). In total, 67,377 genes were annotated, accounting for 90.56% of the genomes (Supplementary file2 (Gene prediction)). The co-annotated gene number is 16,877 for the genomes (Supplementary Fig. S4).

The annotated gene sets were evaluated using BUSCO. Among the 1,614 BUSCO groups, about 98.45% of complete gene elements can be found in the annotated gene set, indicating that the majority of conservative gene predictions are relatively complete and confirming the high reliability of the gene prediction result. Besides, the proportion of expressed genes in four tissues ranged from 71.09–80.34%. Total expressed genes account for 86.60% of the whole gene sets (Supplementary file2 (Transcripts)). The gene structure in *A. donax* genome showed similar distribution trend with other Poaceae plants, including gene length, CDS length, exon length, exon number, intron length and intron number (Supplementary Fig. S5), demonstrating the reliability of the genome annotations.

A. donax is an alloenneaploid

Based on the protein sequence of the genome assembly, we performed intra-genomic comparison within *A. donax*. The discontinuous synteny chromosome segments revealed that *A. donax* undergone multiple chromosome rearrangement during evolutionary process. Interestingly, most single chromosome segment can be aligned to two other chromosome segments (Fig. 3A). Besides, synteny analysis of *A. donax* and *S. italica* showed a 1:3 syntenic relationship (Fig. 3B and Supplementary Fig. S6). Based on these results, we speculated that *A. donax* is an enneaploid.

To determine whether *A. donax* is autoenneaploid or alloenneaploid, we used SubPhaser to split the subgenomes of *A. donax.* The k-mer based heatmap showed that the chromosomes were clustered to two groups, in which subgenome A has 12 chromosomes and subgenome B has 24 chromosomes (Fig. 3C). Therefore, our results jointly proved that *A. donax* is alloenneaploid, and the karyotype is AAABBBBBB (3n = 9x = 108). The whole-genome duplication (WGD) analysis showed that *A. donax* undergone two WGD event, one is the ancient ρ event shared by Poaceae plants (Wang et al. 2015), another is a recent burst WGD event occurred ~ 13.5 MYA (Fig. 3D).

Gene family clustering analysis of *A. donax*

To investigate the genome evolutionary history of *A. donax*, gene family clustering was carried out using *A. donax*, six other Gramineae species (*Oryza sativa*, *Zea mays*, *Sorghum bicolor*, *Brachypodium distachyon*, *Setaria italica* and *Saccharum spontaneum*) and a dicotyledon *Arabidopsis thaliana*. A total of 21,162 gene clusters were identified in *A. donax* genomes, in which 9,063 clusters were shared by all the above species, and 1,989 clusters were unique to *A. donax* (Fig. 4A). A total of 121 Single-copy orthologs shared by *Arundo* and six other grass plants were used for phylogenetic analysis and divergence time estimation, which showed that subfamily *Arundo*, *Setaria and Panicum* shared a common ancestor ~ 49.5 million years ago (MYA) (Fig. 4B).

A. donax undergone dramatic gene family expansion during evolution (Fig. 4C). The 611 expanded gene families were enriched in GO terms like “response to water deprivation”, “response to oxidative stress”, “response to osmotic stress”, “response to cold” and “response to heat” (Fig. 4D), hinting that *A. donax* emerged from the grass family because of the server environment in earth at that time.

Salt stress response gene mining of *A. donax*

Two previous studies had identified multiple salt stress response genes using RNA-seq (Angelo et al. 2019, 2020), while the analysis was based on transcript assembly and offer limited information of specific genes. To deeply mine salt stress response genes, we reanalyzed the public RNA-seq data using the genome assembly above. The mapping rate of RNA-seq data was 67.0%~71.1% and 71.1%~74.4% (Supplementary file 3), furthering proving the accuracy of the assembly. Gene expression heatmap using transcripts per million values (TPM) showed that one of the two studies showed low data consistency among the biological duplications (Supplementary Fig. S7), while another study showed better data

quality (Fig. 5A). Therefore, we used the data with high consistency which contained two gradients of salt treatment (server and extreme) to perform the following analysis.

A total of 3471 differential expression genes (DEGs) were identified in three comparisons. In details, 956 DEGs were identified in CK versus severe (240 up-regulated and 746 down-regulated), 2875 DEGs were identified in CK versus extreme (1119 up-regulated and 1756 down-regulated), 1395 DEGs were identified in severe versus extreme (607 up-regulated and 787 down-regulated) (Fig. 5B and supplementary file 4). Next, 584 DEGs (overlap of CK_severe and CK_extreme) were used for GO enrichment analysis. Interestingly, top 15 enrichment pathways contained “response to water deprivation”, “response to water”, “response to salt”, proving that these DEGs were indeed response to salt stress (Fig. 5C).

Figures

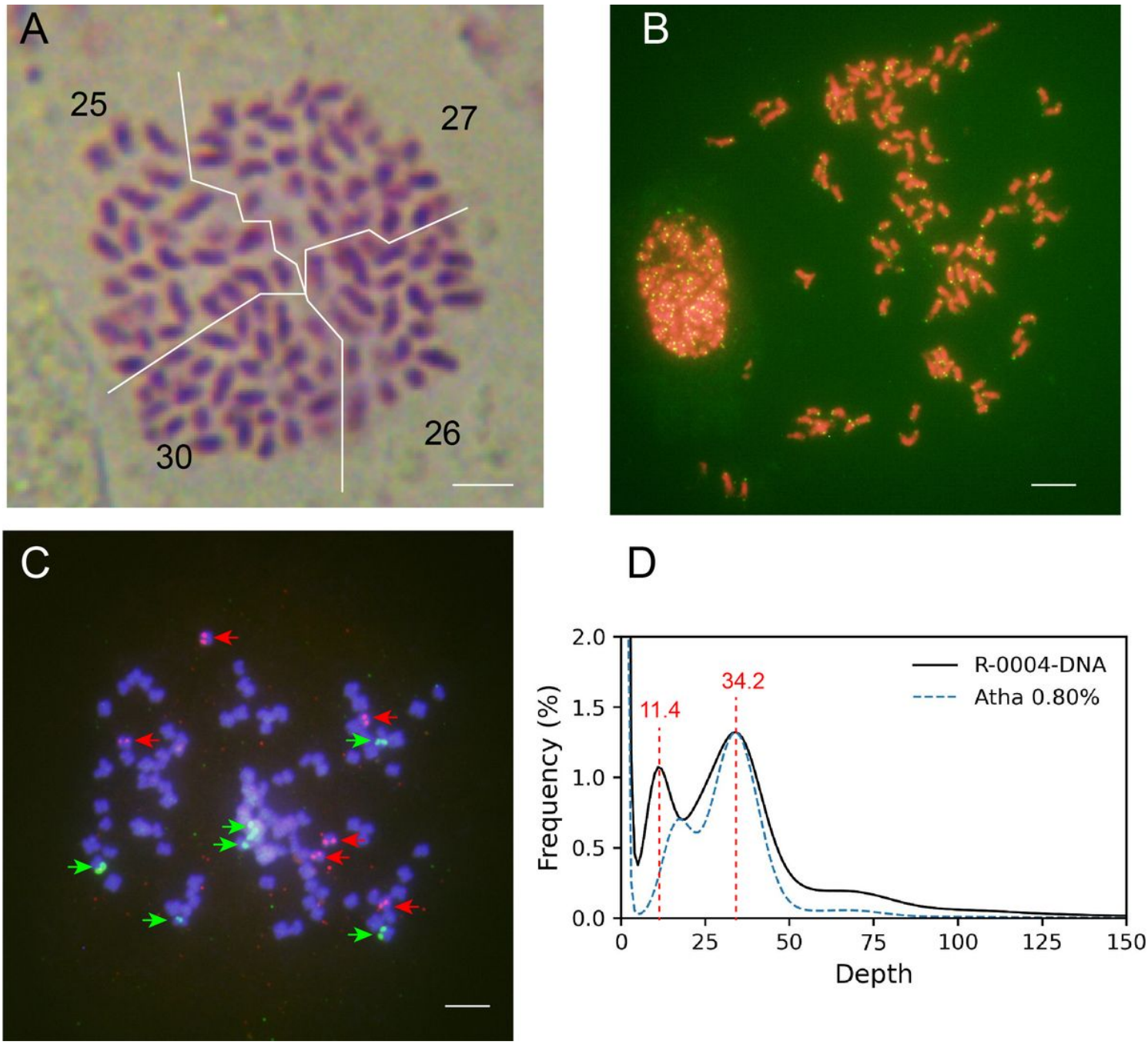

Figure 1

Karyotype and genome survey of *A. donax*. (A) DAPI staining of *A. donax* chromosomes. (B) *In situ* hybridization of telomere repeat sequences. (C) rDNA fluorescence *in situ* hybridization of chromosomes. Red and green arrows denote the 5SrDNA and 18SrDNA hybridization signals. Bar = 50 μm. (D) K-mer distribution curve and heterozygosity simulation curve. Horizontal axis labels the K-mer depth, vertical axis labels frequency of the K-mer depth. Genome of *Arabidopsis thaliana*(Atha) was used to estimate the genome heterozygosity of *A. donax*.

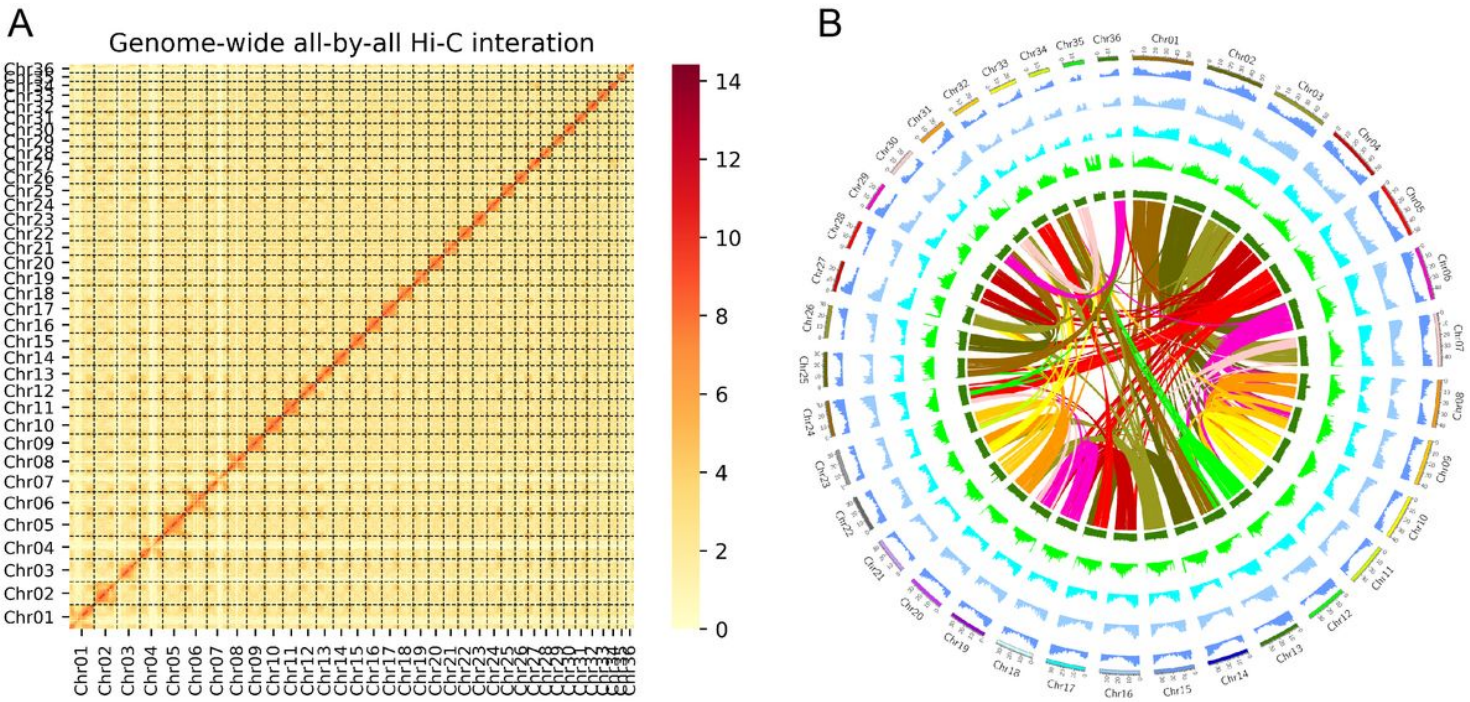

Figure 2

Overview of the *A. donax* genome assembly. (A) Genome-wide Hi-C map of *A. donax*. (B) Circos plots of *A. donax* genome assemblies. The density of GC content (a), repeated sequence distribution (b), gene density (c), functional genes (d), expressed genes (e) were calculated using 100kb non-overlap window. The innermost lines represent inter-chromosomal synteny.

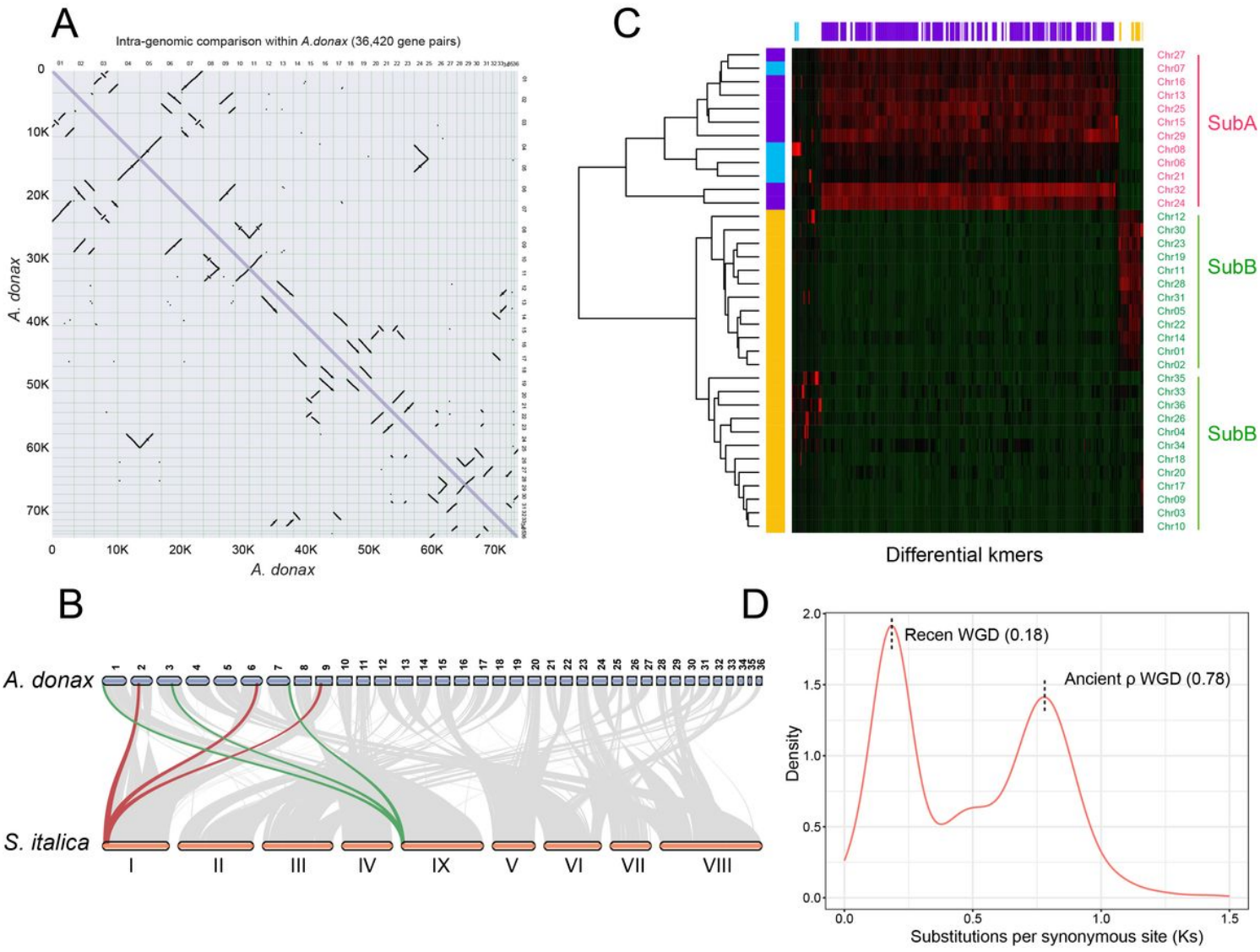

Figure 3

Comparative genomic analysis and whole-genome duplication events of *A. doanx*. Dot plot illustrating the comparative analysis of *A. donax* versus *A. donax*. (B) Syntenic relationships between *A. donax* and *S. italica*. Colored stripes represent the correspondence of syntenic chromosome segments. (C) Subgenomes of *A. doanx*. Clustering based on specific K-mers to partition the two subgenomes. Red and green chromosomes correspond to the A and B subgenomes. (D) The frequency distributions of synonymous substitution rates (Ks) of homologous gene pairs that located in the collinearity blocks of *A. donax* versus *A. donax*. The two peaks predict the recent WGD in *A. donax* and the shared WGD (ρ event) in the grass family.

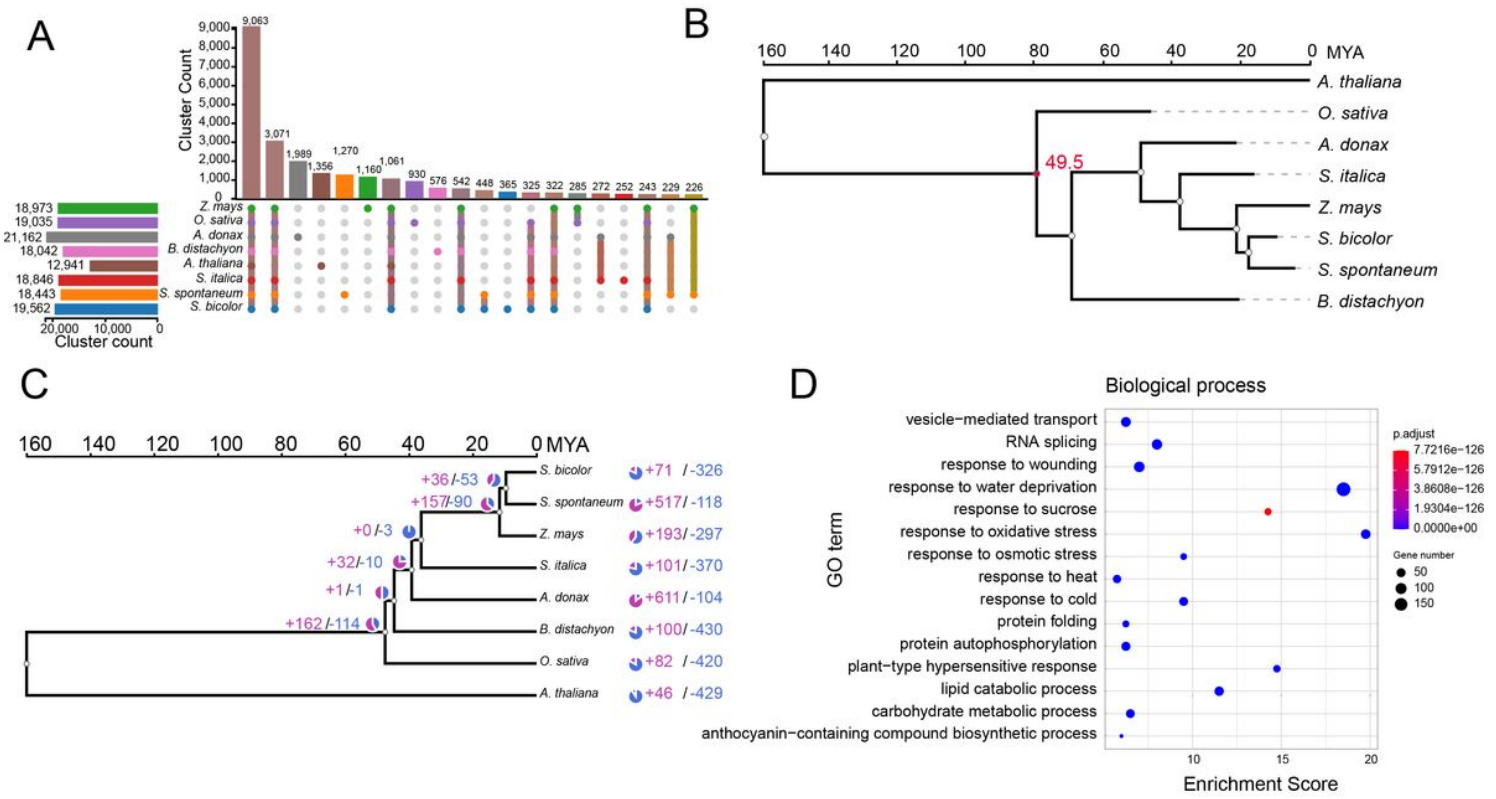

Figure 4

Gene family clustering analysis of *A. donax*. (A) The UpSet table displays the count of orthologous clusters for each species, along with the count of unique gene clusters and the count of shared gene clusters among different species. (B) Timetree of *Arundo* and several other Gramineae plants. The evolutionary relationships between species were estimated by highly conserved single-copy genes. MYA, million years ago. (C) A pie chart was utilized to visualize gene families with altered gene numbers, representing the contracted gene families (depicted in purple) and expanded gene families (depicted in blue). (D) GO enrichment of 611 expanded gene clusters of *A. donax* (top 15 of biological process category). The color of dot denotes the adjusted p-value using hypergeometric test. The size of dot denotes the gene number of GO terms.

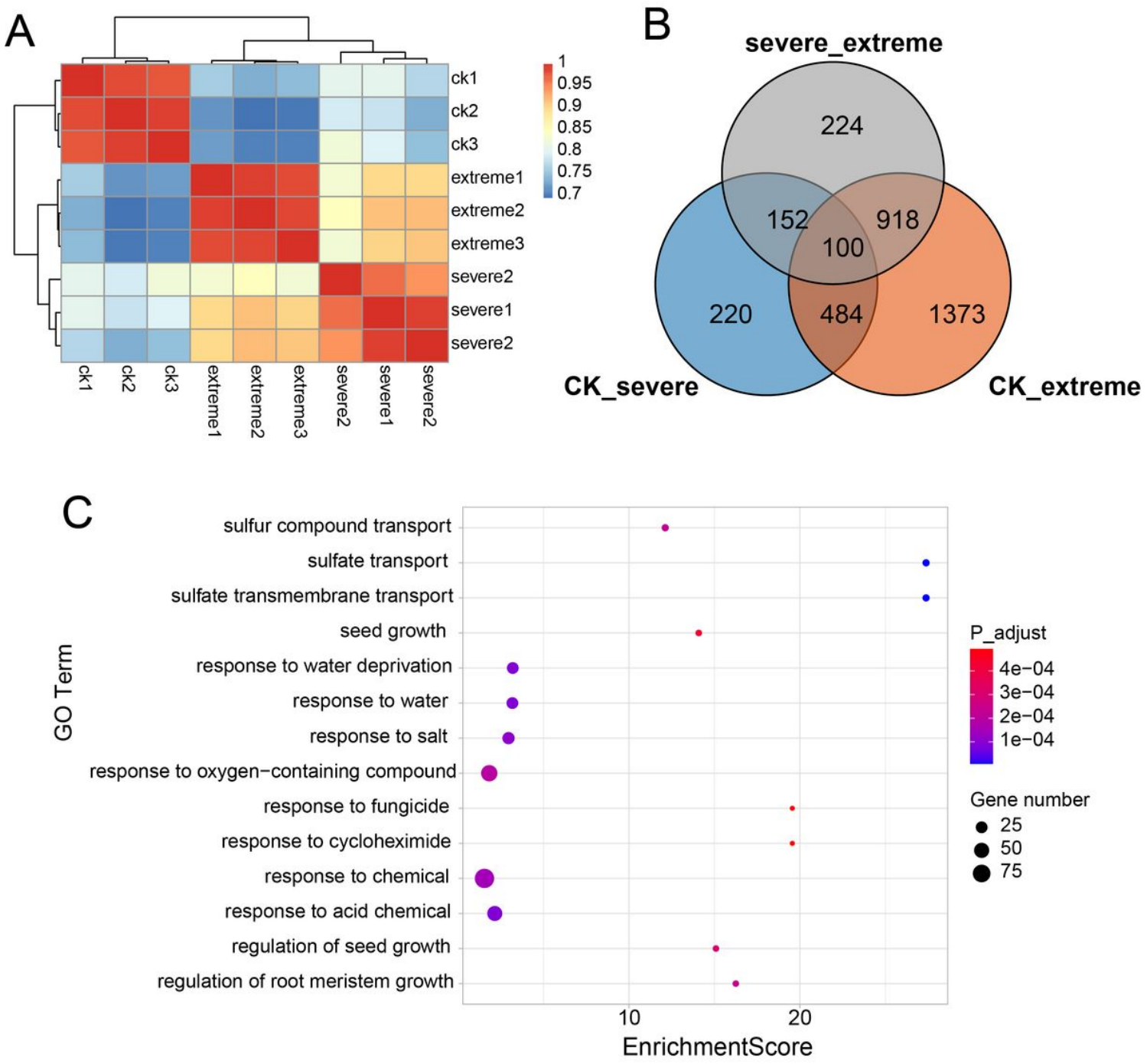

Figure 5

Salt stress response gene mining. (A) Gene expression heatmap of the nine samples using Transcripts Per Million (TPM) values.(B) Veen diagrams of three sets of DEGs. (C) GO enrichment of 584 DEGs shared by CK/server and CK/extreme (top 15 of biological process category). The color of dot denotes the adjusted p-value using hypergeometric test. The size of dot denotes the gene number of GO terms.

Supplementary Files

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

- Supplementaryfile1.docx
- Supplementaryfile2.xlsx
- Supplementaryfile3.xlsx
- Supplementaryfile4.xlsx