一、发表论文情况

（一）1.High-quality chromosome-level de novo assembly of the Trifolium repens（BMC Genomics、第一作者、二区、IF=4.4）

（二）.A High-Quality Genome Assembly of Sorghum dochna（Frontiers in Genetics、第三作者、三区、IF=3.7）

RESEARCH

Open Access

High-quality chromosome-level de novo assembly of the *Trifolium repens*

Hongjie Wang[1,2], Yongqiang Wu[1,2], Yong He[1,2], Guoyu Li[1], Lichao Ma[1,2], Shuo Li[1,2], Jianwei Huang[3] and Guofeng Yang[1,2*]

Abstract

Background White clover (*Trifolium repens L.*), an excellent perennial legume forage, is an allotetraploid native to southeastern Europe and southern Asia. It has high nutritional, ecological, genetic breeding, and medicinal values and exhibits excellent resistance to cold, drought, trample, and weed infestation. Thus, white clover is widely planted in Europe, America, and China; however, the lack of reference genome limits its breeding and cultivation. This study generated a white clover de novo genome assembly at the chromosomal level and annotated its components.

Results The PacBio third-generation Hi-Fi assembly and sequencing methods generated a 1096 Mb genome size of *T. repens*, with contigs of N50 = 14 Mb, scaffolds of N50 = 65 Mb, and BUSCO value of 98.5%. The newly assembled genome has better continuity and integrity than the previously reported white clover reference genome; thus provides important resources for the molecular breeding and evolution of white clover and other forage. Additionally, we annotated 90,128 high-confidence gene models from the genome. White clover was closely related to *Trifolium pratense* and *Trifolium medium* but distantly related to *Glycine max*, *Vigna radiata*, *Medicago truncatula*, and *Cicer arietinum*. The expansion, contraction, and GO functional enrichment analysis of the gene families showed that *T. repens* gene families were associated with biological processes, molecular function, cellular components, and environmental resistance, which explained its excellent agronomic traits.

Conclusions This study reports a high-quality de novo assembly of white clover genome obtained at the chromosomal level using PacBio Hi-Fi sequencing, a third-generation sequencing. The generated high-quality genome assembly of white clover provides a key basis for accelerating the research and molecular breeding of this important forage crop. The genome is also valuable for future studies on legume forage biology, evolution, and genome-wide mapping of quantitative trait loci associated with the relevant agronomic traits.

Keywords *Trifolium repens*, Genome assembly, PacBio HiFi, Genome annotation

Background

White clover (*Trifolium repens L.*) (Fig. 1a), an excellent perennial legume forage, is a heterotetraploid native to southeastern Europe and southern Asia. It is rich in diverse nutrients and mineral elements and has high nutritional, ecological, genetic breeding, and medicinal values [1–4]. The forage also has good palatability for herbivorous livestock, with high carbohydrate and protein content, and is used as ruminant feed in many parts of the world [5, 6]. Moreover, white clover is widely used as lawn ground cover for soil and water conservation due

*Correspondence:
Guofeng Yang
yanggf@qau.edu.cn
[1] College of Grassland Science, Qingdao Agricultural University, Qingdao 266109, China
[2] Key Laboratory of National Forestry and Grassland Administration On Grassland Resources and Ecology in the Yellow River Delta, Qingdao 266109, China
[3] Berry Genomics Corporation, Beijing, China

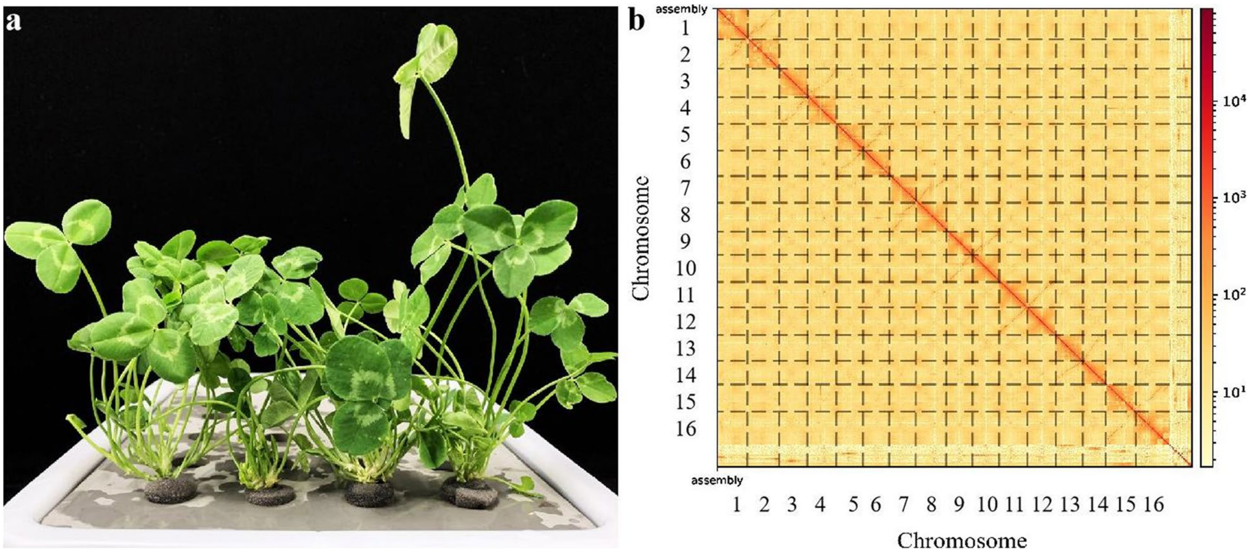


Fig. 1 Plant morphology and Hi-C-assisted genome assembly of white clover (**a**) a Phenotype of the sequenced white clover plant. **b** Hi-C interaction heatmap showing 100-kb resolution super scaffolds

to its soil moisturization effect. White clover exhibits excellent growth when mixed with forages of the family Gramineae. It can play an integral role in intensive grazing systems regarding animal performance and herbage production, thus suggesting its important role in the stable development of the grassland ecosystem [6]. White clover has excellent resistance to cold, drought, trampling, and weed infestation, which is important for improving and breeding new varieties [7–10].

Compared with related species, such as alfalfa and soybean, the structural and genetic information of the white clover is limited, especially at the genomic level, greatly limiting its breeding and improvement [11–13]. Therefore, it is necessary to construct a high-quality white clover genome to accelerate its genetic research and fully use its genetic potential to breed excellent varieties [14].

Here, we use Illumina, PacBio, HiFi, and Hi-C (high-throughput chromatin conformation capture) technologies to generate a high-quality chromosome-level genome assembly of white clover [15, 16]. We annotated the components and functions of the white clover genome and conducted the genomic collinearity analysis between the white clover chromosome and the related species [17]. We also performed the protein family clustering analysis for the predicted genes. Furthermore, phylogenetic trees were constructed to estimate the differentiation time, and the contraction and expansion of gene families on each evolutionary branch were evaluated [18]. Forward selection gene analysis and genome-wide replication analysis were also performed. In summary, this study provides valuable genomic data for further studies and the breeding of white clover. The results of this study also provide a new research direction for analyzing the differentiation and evolution mechanism of white clover and the related species.

Results

Genome-survey, sequencing, and assembly

This study evaluated the size, repetitiveness, heterozygosity, and other genome parameters of the white clover. After quality control, Illumina sequencing yielded 59 Gb of data [19]. Blasting 10,000 randomly selected clean reads against the NT (Nucleotide Sequence Database) library revealed a 98.79% mapping. Moreover, K-mer analysis performed to estimate the complexity of the genome further predicted a genome size of 1075 Mb, with 68.80% repeat and 1.68% heterozygous sequences (Fig. S1). Traditional next-generation sequencing (NGS) data assembly methods were used to predict the genome size, while PacBio HiFi sequencing, a third-generation sequencing (TGS), was conducted for the white clover genome assembly [20]. High-quality Hi-Fi reads were obtained after parameter comparison of the output data. The Hi-Fi reads were 1.89 Mbp in size, with an N50 measure of 16.3kbp.

After eliminating heterozygous and redundant contigs, the assembled genome (1095 Mb) had 380 contigs, with an N50 of 14 Mbp and a maximum contig size of 53 Mbp. The average GC content of the assembled genome was 33.64% (Table 1).

To evaluate the quality and integrity of the assembly, we compared the sequencing data with the assembly

Table 1 Summary statistic for the *Trifolium repens* genome

	Assembly	
Genome assembly	Estimated genome size	1075 Mb
	Total length of assembly	1096 Mb
	Number of contigs	380
	Contig N50	14 Mb
	Largest contig	53 Mb
	Number of scaffolds	202
	scaffold N50	65 Mb
	Chromosome coverage(%)	95.06%
	GC content of genome	33.64%
	Annotation	
		Total length
Transposable elements	Total	672 Mb(61.37%)
	Retrotransposon	448 Mb(40.91%)
	DNA Transposon	140 Mb(12.81%)
		Copies
Noncoding RNAs	rRNAs	10,984
	tRNAs	2,024
	miRNAs	662
	snRNAs	1352
Gene models	Number of genes	90,128
	Mean gene length	3,604 bp
	Mean coding sequence length	1,592 bp

results and found that the mapped ratio was 99.33%, with BUSCO (Benchmarking Universal Single-Copy Orthologs) assembly assessment integrity of 98.50% [21]. The BUSCO results of white clover assembly are shown in Table S1. These results indicate that the assembly had good integrity.

Scaffold construction and curation

In this study, we used the Hi-C technology and generated 270 Gb of data, from which 180 Gb was used to construct chromosome-level super scaffolds with 160 times genome coverage. Subsequent analysis of the Hi-C library revealed a genome with a scaffold-Len of 1096 Mb and an N50 of 65 Mbp. Compared with the previously reported sequence data of white clover (scaffold N50 = 122 kb), the quality and integrity of the data obtained in this study were substantially higher [22].

Moreover, 95.06% of the contigs were attached to 16 chromosomes after the Hi-C-assisted assembly. The genetic material exchange was observed to be much stronger within than between chromosomes [23]. The statistical analysis results of chromosome sequence distribution are shown in Table S2. The heat map showing the genome interaction of the Hi-C-assisted assembly further verified the accuracy of the assembly results (Fig. 1b). Table 1 summarizes the assembly information. Thus, these results demonstrate the high accuracy of the Hi-C assembled genome.

Genome annotation

The gene functions were inferred by analyzing the homology alignments and predicting the repetitive sequences. We constructed a repeat sequence library and annotated 2,023,411 repeat sequences. MITEs (miniature inverted-repeat transposable elements) and LTR (long terminal repeat) transposition components were identified by the structure prediction method, and these elements accounted for 61.37% and 37.75% of the total sequences, respectively. Copia and Gypsy accounted for 13.56% and 11.49% of LTR-retrotransposons, respectively. The results of the repetitive sequences are shown in Table S3.

Additional 4092 simple repeats were also found in the assembled genome, and we predicted 13 types of ncRNA, totaling 15,520 ncRNAs. After removing the gene models containing premature stop codons and frameshifts, we obtained 90,128 high-confidence gene models and 91,690 transcripts using RNA-seq and de novo prediction strategies. However, these gene models were unevenly distributed across the 16 chromosomes.

Each gene contained an average of one transcript, and the average lengths of white clover genes and transcripts were 3604 bp and 1697 bp, respectively. Moreover, each transcript contained an average of 5 exons, with average lengths of 341 bp. We also compared the white clover genome with its five closely related species, including *Medicago truncatula*, *Trifolium medium*, *Vigna radiata*, *Cicer arietinum*, and *Glycine max*. The results showed that *T. medium* (119,102) had the most genes, while *V. radiata* (29,006) and *Cicer arietinum* (28,772) had the least. The five species had similar average coding sequence (CDS) lengths except for *T. medium* (306) (Table 2).

Using the NR, SwissProt, KEGG, GO, and eggNOG databases, we annotated and predicted the function and number of various genes [24]. We annotated 88,094, 61,830, 77,722, 52,992, and 26,979 genes using NR, SwissProt, eggnog, GO, and KEGG databases, respectively. Furthermore, we conducted a Venn analysis by integrating the five databases, which revealed 21,825 common gene annotations (Table S4). Venn analysis of functional gene annotations is shown in Fig. 2.

Gene family and evolution analysis

Closely related species tend to have greater collinear fragments coverage and the collinear relationship between their genomes. Collinearity analysis suggested that the relationship between *T. repens* and *M. truncatula* is relatively close. Moreover, 16 chromosomes of *T. repens* and

Table 2 The information of annotated gene models per species for all the species

Organism	Number of genes	Mean CDS length (bp)	Exons per transcript	Mean exon length (bp)	Mean intron length (bp)
Vigna radiata	29,006	1430	7.6	293	449
Glycine max	54,881	1391	8	295	413
Trifolium medium	119,102	306	1.4	219	172
Cicer arietinum	28,772	1393	7.7	291	418
Medicago truncatula	36,079	1428	6.9	324	393
Trifolium repens	90,128	1592	5	341	490

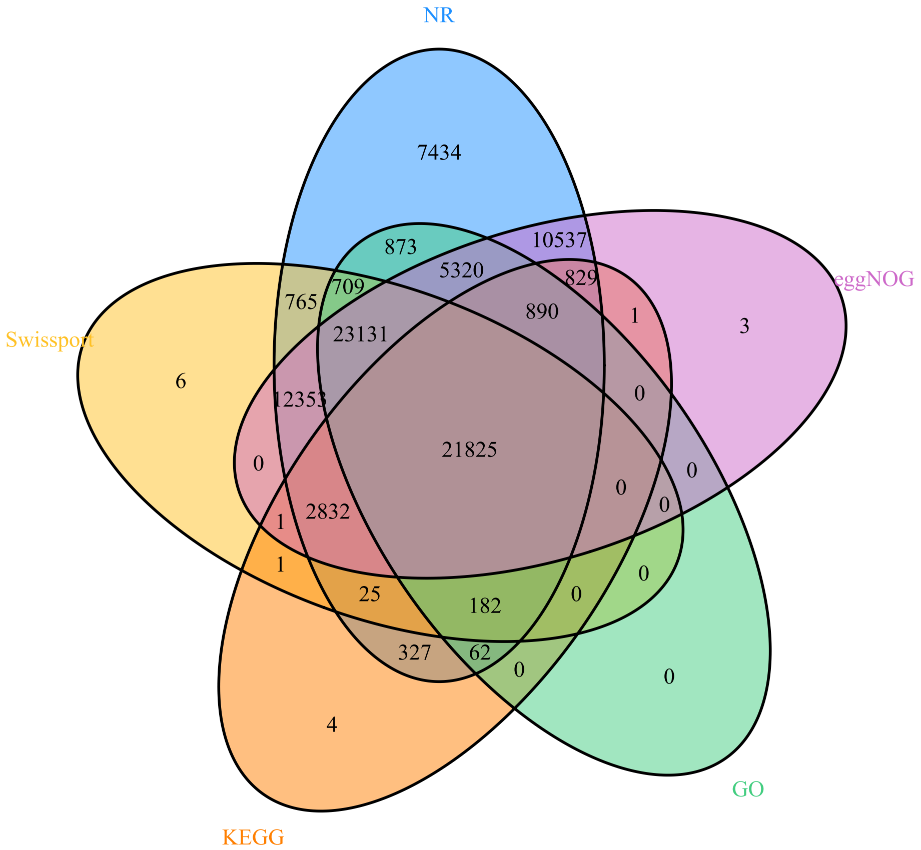


Fig. 2 Venn analysis of five major databases (NR, Swiss-Prot, eggNOG, GO, KEGG) containing gene function annotation information

eight of *M. truncatula* had a good collinear relationship (Fig. 3), indicating their chromosomal conservation after species divergence [25].

The *T. repens* genome assembled in this study was compared with the genomes of seven other related species *G. max*, *V. radiata*, *M. truncatula*, *T. medium*, *C. arietinum*, *Arabidopsis thaliana*, and *T. pratense*. The OrthoMCL clustering analysis showed that 90,128 white clover genes clustered into 25,840 gene families. *Arabidopsis* had the most gene families (26,382), and *T. repens* shared 6,194 gene families with the seven related species (Fig. 4a). Cafe software was used to study the changes in gene families among the species at a family-wide *p*-value threshold of 0.05. The analysis showed that the red trifoliate significantly expanded 1,245 gene families but contracted one gene family during evolution (Fig. 4b) [26]. Distributions of the single-copy genes, multi-copy genes, endemic genes, and other types of genes per species are shown in Supplementary Figure S2.

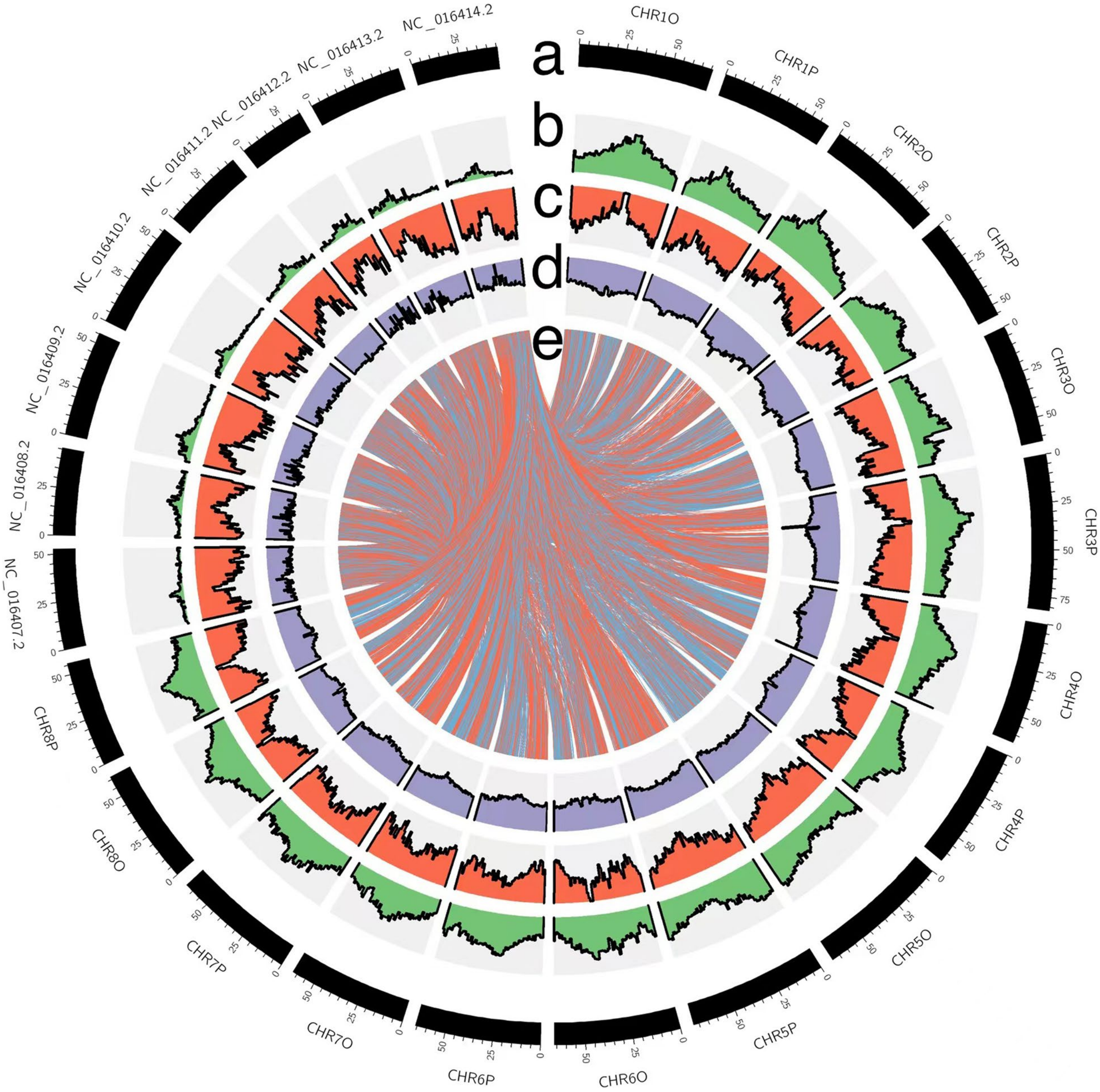


Fig. 3 Features of *T. repens* and *M. truncatula* genome. **a** Length of each pseudochromosome (Mb). **b** Distribution of repetitive sequence. **c** Distribution of gene density. **d** Distribution of the GC content (**e**) *T. repens* and *M. truncatula* synteny analysis; the beginning of NC represents the chromosome of *M. truncatula*, while the beginning of CHR represents the chromosome of *T. repens*

GO functional enrichment analysis revealed the expansion of gene families related to protein phosphorylation, transmembrane transport, adenine nucleotide binding, and membrane composition. Furthermore, *T. repens* gene families were associated with biological processes, molecular function, cellular components, and environmental resistance, which could explain its excellent agronomic traits (Table S5). The positive selection analysis model was established with white clover as the foreground branch and other species as the background branch. Finally, three genes with significant positive selection were obtained.

We constructed a phylogenetic tree based on the results of protein family clustering and found that *T. repens* formed a monophyletic group with *V. radiata*, *G. max*, *T. pratense*, *T. medium*, *M. truncatula*, and *C. arietinum* [27]. White clover was most closely related

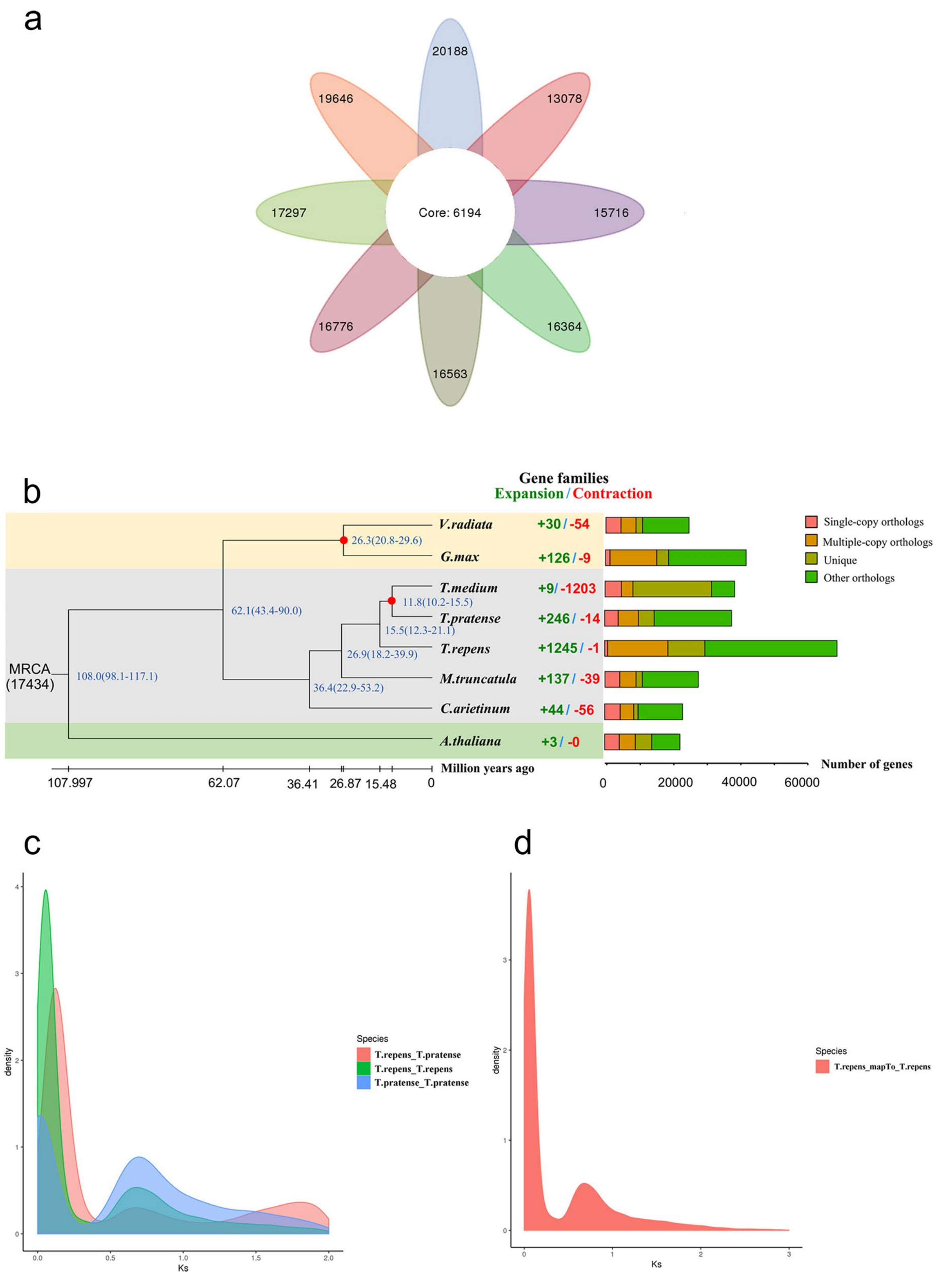


Fig. 4 Gene family and phylogenetic tree analyses of white clover and other representative plant genomes. **a** Venn diagram of the number of shared gene families. **b** A phylogenetic tree based on shared single-copy gene families (left), gene family expansions and contractions among white clover and seven other species (middle), and Gene family clustering in white clover and seven other plant genomes (right). **c** Genome-wide replication Ks distribution map of white clover and its related species. **d** Genome-wide replication Ks analysis of white clover

to *T. pratense* and *T. medium*, with their estimated divergence time being 15.5 million years ago (Fig. 4b).

Whole genome duplication (WGD) events are important indices of plant evolution and are the driving force for plant adaptation to various environments [28]. Thus, WGD provides sufficient genetic material for expanding plant gene families or generating new genes. It also enhances the adaptability of plants to the environment and accelerates the evolution of plants by generating various genetic variations. To explore the evolutionary history of *T. repens*, we used the changes in the synonymous replacement rate of paralogous genes to measure gene duplication and loss in its genome. The resultant data suggested that the divergence of *T. repens* and *T. pratense* occurred after the WGD events. Both *T. repens* and *T. pratense* experienced a WGD event when the K_S value was 0.13 (Fig. 4c); however, an additional WGD event also occurred when the K_S value of *T. repens* was 0.6 (Fig. 4d).

Discussion

Leguminous forages have excellent agronomic traits, and their genomic data are important for genetic analysis, breeding, and functional omics. White clover is a forage and lawn grass widely grown worldwide. Assembling white clover (*T. repens*) is challenging due to its large genome structure and highly homologous genomic sequences. However, this study assembled a high-quality tetraploid white clover genome using the latest third-generation Hi-Fi assembly and sequencing methods, providing a good reference for the research on other herbage of the Clover genus.

Compared with the second-generation sequencing technology, TGS technology overcomes some NGS shortcomings in genome assembly. TGS does not require polymerase chain reaction (PCR) amplification or long read length and has no guanine-cytosine (GC) preference, thus making genome assembly using PacBio Hi-Fi an effective assembly strategy [29, 30]. Compared with the previously published 841 Mb (N50 = 122 kb) genome assembly of white clover, the genome size in this study was 1,095 Mb (contig N50 = 14 Mbp), indicating a significantly improved quality. The average GC content of the assembled genome was 33.64% (Table 1), close to the previously assembled *Trifolium repens* genome (35%). Additionally, the number of newly assembled white clover reads was 400,467,170, and the mapping ratio was 99.33%, higher than that of the previously assembled genome (98%). Compared with the previously reported total BUSCO groups (1321), the assembly in this study had 2326 total BUSCO groups, and the Fragmented and Missing BUSCOs were smaller than the previous assembly. Moreover, complete Single-Copy BUSCOs (98.5%) were higher in the present than in the previous assembly (92%). Thus, the newly assembled white clover genome had better continuity and integrity than its previously reported reference genome [22].

Thus, our work has provided a chromosomal-level genome assembly using Hi-C-assisted genome assembly of white clover based on the whole-genome data. The technique utilizes the entire cell nucleus to fix and capture the mutual chromosomal sites [31, 32]. Hi-C uses high-throughput sequencing to determine the whole-genome spatial distribution of chromatin DNA through a high-resolution interaction map of chromatin regulatory elements obtained from the positional relationship [33, 34]. The published examples of higher plants assembled with Hi-C-assisted genomes include quinoa, barley, durian, and so on [32, 35, 36]. In this study, the assembly generated contains 202 scaffolds (~1096 Mb) spanning N50 = 65 Mb, with significantly improved quality. The assembled genome had higher coverage (95.06%) at the chromosomal level after high-throughput sequencing and Hi-C scaffolding.

We annotated 90,128 high-confidence gene models from the newly assembled genome. The published assembled genome annotated 68,558 genes, and the average CDS length was larger than the reported genome [22]. A high-quality reference genome of *T. repens* is important for understanding its evolution, origin, and domestication history. Therefore, this study provides important resources for molecular breeding and evolution analysis of white clover and other forages [37].

T. occidentale and *T. pallescens* are reportedly the progenitors of white clover, which originated about 15–28,000 years ago from multiple hybridization events during the last glaciation. Therefore, its evolutionary history is not well-understood. Genomic collinearity analysis showed that *T. repens* and *M. truncatula* exhibited close phylogenetic and genetic relationships. Moreover, phylogenetic analyses revealed that *T. repens* diverged after *V. radiata*, *G. max*, *M. truncatula*, and *C. arietinum* but before *T. medium* and *T. pratense* [38]. Thus, these species share the same ancestry with *T. repens*.

This study focused on comparing the genomes of white clover and related species at the genomic level. The structural genome characteristics, gene function, and evolutionary status of white clover were explained by the collinearity analysis between related species and intraspecies. Moreover, whole-genome replication events, phylogenetic tree construction, and differentiation time estimation, gene protein family clustering, contraction/expansion analysis, gene retention and loss, and forward selection gene analysis were also conducted. In summary, we decoded the complex white clover genome, revealed the events that have shaped the genome, and created foundations for further studies on legumes and complex genome assembly [20, 38]. The newly assembled genome is also valuable for future studies on white clover biology, evolution, and genome-wide mapping of quantitative trait loci associated with its agronomic traits.

Future research on this work will focus on the in-depth evaluation of specific traits of white clover, transcriptome sequencing, or large-scale population resequencing of specific tissue sites or growth and development periods. We will also consider using high-resolution single-cell technology to conduct single-cell transcriptome analysis of specific tissue sites in an attempt to solve the molecular mechanism of white clover resistance to various stresses. This will provide a valuable reference for further studies and utilization of white clover, an important forage resource.

Conclusions

This study reported a high-quality de novo assembly for white clover obtained at the chromosomal level using PacBio third-generation Hi-Fi sequencing. The newly assembled genome has outstanding coverage and integrity; thus provides a key basis for accelerating the research and molecular breeding of this important forage crop. The genome is also valuable for future studies on white clover biology, evolution, and genome-wide mapping of quantitative trait loci associated with its agronomic traits.

Experimental procedures

The *T. repens* (2n = 4x = 32) was planted in a light incubator at the Key Laboratory of National Forestry and Grassland Administration on Grassland Resources and Ecology in the Yellow River Delta. Thereafter, five-week-old leaf samples were sampled from each white clover into vacutainer tubes for genomic DNA extraction. The study complied with the ethical norms of Chinese and international regulations.

DNA and RNA extraction

The *T. repens* (white clover Super Haifa) plants were grown in a phytotron chamber at 25 °C at the Qingdao Agricultural University in Shandong, China, under the photoperiod of 16/8 h, a light intensity of 400 W/m ^{2}, and relative humidity (RH) of 70%. The leaf samples were collected and treated with liquid nitrogen for DNA extraction using the Tiangen DNA Secure Kit for Genome Sequencing (Beijing, China). Total RNA was extracted using an EASYspin Plus Polysaccharide Polyphenols/Complex Plant RNA Rapid Extraction Kit, following the manufacturer's instructions.

Survey analysis

The quality and quantity of DNA samples were controlled, and the qualified DNA samples were randomly broken into fragments by Covaris ultrasonic fragmentation instrument. Library preparation was conducted by terminal repair, a-tail addition, sequencing connector addition, purification, and PCR amplification. The libraries were then subjected to paired-end 150 (PE150) sequencing using Illumina NovaSeq [39–41]. The original image data file sequenced by the high-throughput sequencer was converted into the original sequence by base calling analysis [39]. To obtain clean reads, we filtered the off-plane reads to remove joints with low numbers, repeated and low-quality reads that would affect the comparison quality and subsequent analysis. We randomly selected and blasted 10,000 clean reads against the NCBI non-redundant nucleotide database (NT library) to check for possible external contamination [42].

K-mer analysis using Jellyfish software estimated the genome size, sample heterozygosity, and genome repeat sequence ratio (Table S6) [43]. The genome size of white clover was estimated using the following formula: G = Knum/Kdepth, where Knum is the number of k-mers, while Kdepth is the expected depth of k-mers.

Genome assembly and quality evaluation

Minia was used for preliminary assembly with second-generation data (Table S6), and the assembly results were evaluated using the GC_depth analysis. DNA concentration and purity were measured by NanoDrop 2000 spectrophotometry. After sequencing with PacBio SMRT technology, a PCR-free SMRTbell library was constructed from a high-quality purified genome through repair and end-joining [44]. The library size was then determined by pulsed-field electrophoresis, and the acquired data were filtered and loaded onto smrtlink (Table S6) for CCS (Circular Consensus Sequencing) processing. The original PacBio Sequel data, the Polymerase Reads, were filtered to obtain subsequent available SubReads, which were then processed with smrtlink software for CCS to obtain high-quality HiFi Reads (Table S6). To obtain high-quality Hi-Fi Reads, we conducted CCS on the SubReads obtained above using parameters –min-passes = 3 –min-rq = 0.99.

Hifiasm software (Table S6) was used for assembly, and all-vs-all alignment was used to correct sequencing errors [20, 44, 45]. Overlap comparison was repeated after correction, and a phased string graph was constructed [41]. Finally, contigs were generated based on the overlapping graph, and the assembly contig sequence was further deheterozygosed through purge_dups(v1.2.3) (Table S6) [46, 47]. The assembled genome was compared with HiFi Reads using the software Minimap2 (Table S6), and then heterozygotic fragments were removed based on the coverage distribution and sequence score of the reads [48]. Pseudo contigs is removed from the genome by BWA (Burrows-Wheeler-Alignment Tool) (Table S6). After redundancy analysis, the genome sequence was compared with the second-generation data, and the GC-depth graph was generated. Contigs with average coverage depth less than 5X was removed. In addition, contigs with window GC content of 50%-53% were also removed, and the final assembly result was calculated.

The sequencing data was compared with the assembly results to evaluate the data recovery ratio and integrity assessment was conducted using BUSCO (Table S6) and the BUSCO Eudicots lineage dataset (eudicots_odb10) [21]. The genome assembly results were evaluated based on the proportion of matched read pairs and the distribution of inserted fragments. Tblastn (Table S6), Augustus, and Hmmer tools were used to evaluate the integrity of the single-copy orthologous genes [21]. Genome sequencing was performed by Berry Hekang (Beijing, China) using the third-generation PacBio Sequel II sequencing platform.

Hi-C data analysis and chromosome construction

For DNA cross-linking, we soaked 100 mg of *T. repens* leaf tissues in paraformaldehyde (a cell cross-linking agent) for 15 min, after which glycine was added to terminate the chromatin cross-linking reaction. The treated tissues were collected, frozen in liquid nitrogen and ground for DNA extraction. Biotin-labeled oligonucleotide ends were added during the terminal repair, and the adjacent DNA fragments were linked with nucleic acid ligase. The protein was enzymatically cleaved at the junction point with protease, and the Covaris crusher was used to randomly break up 350 bp of DNA [35, 49]. Biotinylated DNA fragments were bound to avidin magnetic beads to create the whole library. After qualified library analysis, different libraries were pooled for Illumina PE150 sequencing according to the concentration and target requirements for machine data volume [19]. Thereafter, 10,000 pairs of sequencing reads were randomly selected from the Hi-C sequencing database and blasted against the NT library (Table S6). The top 10 matched species were sequenced and evaluated to determine whether there was bacterial contamination. The JUICER (Table S6) software was then employed to compare the Hi-C data with the draft genome [31, 45, 50]. We analyzed the Hi-C library results via 3D-DNA (Table S6) comparison to obtain valid Hi-C data and generate the chromosome-level scaffold of the white clover genome [31, 45]. After the Hi-C-assisted assembly was completed, the interchromosome and intra-chromosome exchanges were calculated to further verify the accuracy of the assembly results [19].

Genome annotation

Repetitive sequences of the white clover genome were annotated using homology-based and ab initio search methods [51, 52]. Class II transposition factor mites and involuntary transposition factors less than 2 kb in length were searched in the genome using MITEs [53]. To obtain more reliable LTR-RT, we used an LTR retriever to analyze the process. We combined LTRharvest (-similar 90 -vic 10 -seed 20 -seqids yes -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1) with LTR Finder results to filter false-positive LTR-RT (Table S6) [54, 55]. Repetitive sequences of known species were searched in the RepBase library using RepeatMasker (http://www.girinst.org/server/RepBase/index.php) in combination with MITEs.lib library and Lcr.lib library. The combination library was then used as the database to shield the repetitive sequences of the genome using RepeatMasker (Table S6), which were re-identified using RepeatModeler (Table S6). The sequences classified as unknown by the RepeatModeler were compared with the transposable enzyme database using Blastx, and reclassified according to the transposable enzyme type.

The tRNA ab initio rRNA was predicted using tRNAscan-SE (Table S6) software [56], and the other types of ncRNA were searched using the Rfam database (ftp://ftp.ebi.ac.uk/pub/databases/Rfam/14.1/) [56–59]. The specific information of these RNA types was obtained through similarity comparison.

All repetitive regions except the tandem repeats were soft-masked for protein-coding gene annotation. The coding sequences of *M. truncatula* (GCF 000219495.3 MedtrA17 4.0), *T. medium* (GCA 003490085.1 ASM349008v1), *V. radiata* (GCF 000741045.1 *Vradiata* ver6), *C. arietinum* (GCF 000331145.1 ASM33114v1), and *G. max* (GCF 000004515.5 *Glycine max* v2.1) were downloaded. These coding sequences were then subjected to blast (Table S6) searches against the white clover genome, and the homologs containing premature stop codons and frameshifts were discarded [42]. GeMoMa-1.6.1 (Table S6) was used to compare the protein sequence of the related species with the assembled genome to predict their gene structure. Meanwhile, the boundary information of exon and intron was obtained by comparing RNA data with the assembly results. High-quality full-length transcripts were established through the iso-seq standardization process in SMRT analysis software and used to predict the open reading frames (ORFs) via PASA v2.0.1 (Table S6). The protein sequences were filtered to 100AA ~ 1000AA and a CDS number of ≥ 2. A gene that matched the full length of the reference protein sequence was obtained, and the cDNA sequence of the gene was used as the training set. Augustus, SNAP, GlimmerHMM, and GeneMark-ESSuite (Table S6) were used to predict the gene structure [60]. The training set was used for parameter training, and the intron hints indicated that the RNA-Seq reads and scaffolds were comparable. The compared reads were then combined with intron hints for gene structure prediction. The predictions obtained using these packages were combined using EVM (Table S6), after which 36,511 genes were retrieved and functionally annotated by blast searches against NR (ftp://ftp.ncbi.nlm.nih.gov/blast/db/), Swiss-Prot (ftp://ftp.ebi.ac.uk/pub/databases/uniprot/knowledgebase/uniprot_sprot.fasta.gz), eggNOG, GO (http://geneontology.org/), and KEGG (http://www.genome.jp/kegg/) databases. Venn analysis of these databases was then performed to obtain more accurate gene functional annotation information [61].

Genome comparative analysis

We conducted genome collinearity analysis of the white clover and its relatives using the Mummer software (parameters: nucmer -g 1000 -c 90 -l 200) and Lastz (Table S6) [62, 63]. To determine the similarity between sequences, we used OrthoMCL (Table S6) clustering analysis to perform all-VS-All BLAST alignment on gene protein-coding sequences of all selected species (e-value = 1e-5 by default) [64]. Markov clustering algorithm was used for clustering analysis (expansion coefficient is 1.5), and the clustering results distinguished between the endemic and common genes, as depicted by the Venn diagram [64, 65].

The Mafft (Table S6) software was subsequently used for multiple sequence comparisons of supergenes [66]. A suitable base substitution model was selected, followed by constructing a species-based maximum likelihood (ML) phylogenetic tree [27, 67, 68]. Moreover, the mcmctree tool of the PAML (Table S6) software package (parameters: burn-in = 5,000,000, sample-number = 1,000,000, sample-frequency = 50) was used to estimate the differentiation time based on the single-copy gene family [69, 70]. The gene families of each species were then analyzed using the Café (Table S6) software. The numbers of gene family contractions and expansions on each evolutionary branch were obtained, and their occurrences were assessed. After the threshold value of the family-wide *P*-value was set at 0.05, GO functional enrichment analysis was performed for genes in these families.

Furthermore, protein-coding sequences were identified using the positive selection approach by distinguishing between synonymous substitutions (Ks) and non-synonymous substitutions (Ka) [71]. The analysis method of the Branch-site model proposed in 2002 can detect the forward selection occurring in a specific evolutionary lineage and affecting only a portion of genome sites [72]. This study used the Branch-site model to detect the forward selection acting on the protein-coding sequence. Briefly, one-to-one orthology proteins from white clover and related species were selected, and homologous protein sequences were compared using the default parameters of PRANK. The alignment results were filtered with Gblocks (parameters: -t = c -e = .ft -b4 = 5 -d = y), and CODEML in PAML was used to test the positive selection in a specific branch, which only affected some loci. Thereafter, the Chi2 program in PAML (Table S6) was used to check and correct multiple hypotheses (Main parameters include; degree of freedom = 2), after which we obtained the positive selection genes.

Ks values for homoeologous loci of the constructed genome were used to detect WGD events [73]. Moreover, Blastp was used to compare the longest protein sequence encoded by the white clover genes. The MCScanX (Table S6) software was subsequently used to filter the comparison results, and the Yn00 tool of the PAML (Table S6) software package was used to calculate the synonymous replacement rate [74, 75]. Furthermore, a density distribution map based on the Ks values of all paralog and ortholog gene pairs between the genomes of white clover, red clover, and other related species was drawn using MATLAB [26, 76]. The gene comparisons were then made between and within related species.

Abbreviations

NT	Nucleotide Sequence Database
PE	Paired-end
NGS	Next-Generation Sequencing
CCS	Circular Consensus Sequencing
BUSCO	Benchmarking Universal Single-Copy Orthologs
Hi-C	High-throughput chromosome conformation capture
MITEs	Miniature inverted repeat transposable elements
LTR	Long terminal repeat
LTR-RT	Long terminal repeat retrotransposons
ncRNA	Non-coding RNA
NR	NR is the NCBI non-redundant protein database
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
WGD	Whole Genome Duplications
BWA	Burrows-Wheeler-Alignment Tool

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09437-8.

Additional file 1.

Additional file 2.

Additional file 3: Table S1. Benchmarking Universal Single-Copy Orthologs analysis of white clover.

Additional file 4: Table S2. Chromosome sequence distribution statistics.

Additional file 5: Table S3. Repeat sequences results.

Additional file 6: Table S4. GO, eggNOG, NR, KEGG and SP annotation results.

Additional file 7: Table S5. The gene families described (including their GO terms) and their numbers between white clover and the gene family expansion in white clover.

Additional file 8: Table S6. URLs and code of the software.

Acknowledgements

The author would like to thank Professor Guofeng Yang, Professor Zengyu Wang, and Professor Juan Sun (Professor of Grassland Science, Qingdao Agricultural University) for their help in data analysis and article writing. Thank you for the scientific research funding provided by the College of Grassland Science of Qingdao Agricultural University. Thanks for the experimental help provided by Beijing Berry and Kang.

Authors' contributions

HW and GY conceived and designed this research. HW analyzed data and wrote the manuscript. HW, YW, YH and GL executed the data analyses. LM participated in the discussionof the results. YW, YH, LM, and SL collected samples. GY, SL, JH contributed to the evaluation and discussion of the results and manuscript revisions. All authors have read and approved the final version.

Funding

This study was supported by the National Nature Science Foundation of China (U1906201), Shandong Forage Research System (SDAIT-23–01), China Agriculture Research System (CARS-34) and the First Class Grassland Science Discipline Program of Shandong Province (1619002), China.

Availability of data and materials

All data generated and analyzed during this current study are available in the Grassland Agri-husbandry Research Center, Qingdao Agricultural University with permission from the Competent Authority. All raw data data were submitted in NCBI Database (SAMN22208873, SAMN33387310, SRR16288262) and the genome assembly and annotation were uploaded in the dedicated public repositories (De novo assembly of Trifolium repens: 10.6084/m9.figshare.23266319, genome annotation of Trifolium repens: 10.6084/m9.figshare.23266532). The details of software used are in Table S6. Biological materials used in this study available from the corresponding author.

Declarations

Ethics approval and consent to participate

T. repens is not endangered or a protected species in China, and it was purchased from BEST grass industry and planted in a light incubator. The seeds are collected by Professor Guofeng Yang in BEST grass industry. All the study procedures were carried out in accordance with relevant guidelines.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 27 February 2023 Accepted: 8 June 2023
Published online: 13 June 2023

References

1. Vrignon-Brenas S, Celette F, Piquet-Pissaloux A, Corre-Hellou G, David C. Intercropping strategies of white clover with organic wheat to improve the trade-off between wheat yield, protein content and the provision of ecological services by white clover. Field Crop Res. 2018;224:160–9.
2. Guy C, Hennessy D, Gilliland TJ, Coughlan F, McClearn B, Dineen M, McCarthy B. White clover incorporation at high nitrogen application levels: results from a 3-year study. Anim Prod Sci. 2020;60(1):187–91.
3. Sabudak T, Guler N, Trifolium L. --a review on its phytochemical and pharmacological profile. Phytother Res : PTR. 2009;23(3):439–46.
4. Chen Y, Chen P, Wang Y, Yang C, Wu X, Wu C, Luo L, Wang Q, Niu C, Yao J. Structural characterization and anti-inflammatory activity evaluation of chemical constituents in the extract of Trifolium repens L. J Food Biochem. 2019;43(9): e12981.
5. Deguchi S, Uozumi S, Touno E, Uchino H, Kaneko M, Tawaraya K. White clover living mulch reduces the need for phosphorus fertilizer application to corn. Eur J Agron. 2017;86:87–92.
6. Egan M, Galvin N, Hennessy D. Incorporating white clover (Trifolium repens L.) into perennial ryegrass (Lolium perenne L.) swards receiving varying levels of nitrogen fertilizer: Effects on milk and herbage production. J Dairy Sci. 2018;101(4):3412–27.
7. Zhang XQ, Yang HH, Li MM, Chen C, Bai Y, Guo DL, Guo CH, Shu YJ. Time-course RNA-seq analysis provides an improved understanding of genetic regulation in response to cold stress from white clover (Trifolium repens L.). Biotechnol Biotec Eq. 2022;36(1):745–52.
8. Nichols SN, Hofmann RW, Williams WM. Drought resistance of Trifolium repens x Trifolium uniflorum interspecific hybrids. Crop Pasture Sci. 2014;65(9):911–21.
9. Ludvikova V, Pavlu VV, Gaisler J, Hejcman M, Pavlu L. Long term defoliation by cattle grazing with and without trampling differently affects soil penetration resistance and plant species composition in Agrostis capillaris grassland. Agr Ecosyst Environ. 2014;197:204–11.
10. Vrignon-Brenas S, Celette F, Amosse C, David C. Effect of spring fertilization on ecosystem services of organic wheat and clover relay intercrops. Eur J Agron. 2016;73:73–82.
11. Chakrabarti M, Dinkins R, Hunt A: De novo transcriptome assembly and dynamic spatial gene expression analysis in red clover. The Plant Genome 2016;9(2).
12. Chen H, Zeng Y, Yang Y, Huang L, Tang B, Zhang H, Hao F, Liu W, Li Y, Liu Y, et al. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nat Commun. 2020;11(1):2494.
13. Wang T, Ren L, Li C, Zhang D, Zhang X, Zhou G, Gao D, Chen R, Chen Y, Wang Z, et al. The genome of a wild Medicago species provides insights into the tolerant mechanisms of legume forage to environmental stress. Bmc Biol. 2021;19(1):96.
14. Kuon J, Qi W, Schläpfer P, Hirsch-Hoffmann M, von Bieberstein P, Patrignani A, Poveda L, Grob S, Keller M, Shimizu-Inatsugi R, et al. Haplotype-resolved genomes of geminivirus-resistant and geminivirus-susceptible African cassava cultivars. Bmc Biol. 2019;17(1):75.
15. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356(6333):92–5.
16. Koren S, Walenz B, Berlin K, Miller J, Bergman N, Phillippy A. kCanu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.
17. Cui FC, Taier G, Li ML, Dai XX, Hang N, Zhang XZ, Wang XF, Wang KH. The genome of the warm-season turfgrass African bermudagrass (Cynodon transvaalensis). Hortic Res-England. 2021;8(1):16.
18. Hubner S, Bercovich N, Todesco M, Mandel JR, Odenheimer J, Ziegler E, Lee JS, Baute GJ, Owens GL, Grassa CJ, et al. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Nat Plants. 2019;5(1):54–62.
19. Dudchenko O, Batra S, Omer A, Nyquist S, Hoeger M, Durand N, Shamim M, Machol I, Lander E, Aiden A, et al. Aedes aegyptiDe novo assembly of the genome using Hi-C yields chromosome-length scaffolds. Science (New York, NY). 2017;356(6333):92–5.
20. Cheng H, Concepcion G, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170–5.
21. Seppey M, Manni M, Zdobnov E. BUSCO: assessing genome assembly and annotation completeness. Methods Mole Biol (Clifton, NJ). 2019;1962:227–45.
22. Griffiths A, Moraga R, Tausen M, Gupta V, Bilton T, Campbell M, Ashby R, Nagy I, Khan A, Larking A, et al. Breaking free: the genomics of allopolyploidy-facilitated niche expansion in white clover. Plant Cell. 2019;31(7):1466–87.

23. Maughan P, Lee R, Walstead R, Vickerstaff R, Fogarty M, Brouwer C, Reid R, Jay J, Bekele W, Jackson E, et al. Genomic insights from the first chromosome-scale assemblies of oat (Avena spp.) diploid species. Bmc Biol. 2019;17(1):92.
24. Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinf. 2013;10(3):645–56.
25. Shen C, Du H, Chen Z, Lu H, Zhu F, Chen H, Meng X, Liu Q, Liu P, Zheng L, et al. The chromosome-level genome sequence of the Autotetraploid Alfalfa and resequencing of core germplasms provide genomic resources for Alfalfa research. Mol Plant. 2020;13(9):1250–61.
26. Hahn M, De Bie T, Stajich J, Nguyen C, Cristianini N. Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res. 2005;15(8):1153–60.
27. Vanneste K, Van de Peer Y, Maere S. Inference of genome duplications from age distributions revisited. Mol Biol Evol. 2013;30(1):177–90.
28. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Da Silva C, Labadie K, Alberti A, et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun. 2014;5:3657.
29. Athanasopoulou K, Boti M, Adamopoulos P, Skourou P, Scorilas A. Third-generation sequencing: the spearhead towards the radical transformation of modern genomics. Life (Basel, Switzerland). 2021;12(1):30.
30. Hassan S, Bahar R, Johan M, Mohamed Hashim E, Abdullah W, Esa E, Abdul Hamid F, Zulkafli Z. Next-Generation Sequencing (NGS) and Third-Generation Sequencing (TGS) for the Diagnosis of Thalassemia. Diagnostics (Basel, Switzerland). 2023;13(3):373.
31. Durand N, Shamim M, Machol I, Rao S, Huntley M, Lander E, Aiden E. juicer provides a one-click system for analyzing loop-resolution hi-C experiments. Cell Syst. 2016;3(1):95–8.
32. Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, Rajasegaran V, Lim WK, Ong CK, Chan K, Cheng VKY, et al. The draft genome of tropical fruit durian (Durio zibethinus). Nature Genet. 2017;49(11):1633-+.
33. Kong S, Zhang Y. Deciphering hi-C: from 3D genome to function. Cell Biol Toxicol. 2019;35(1):15–32.
34. Eagen K. Principles of chromosome architecture revealed by hi-C. Trends Biochem Sci. 2018;43(6):469–78.
35. Jarvis DE, Ho YS, Lightfoot DJ, Schmockel SM, Li B, Borm TJA, Ohyanagi H, Mineta K, Michell CT, Saber N, et al. The genome of Chenopodium quinoa (vol 542, pg 307, 2017). Nature. 2017;545(7655):510–510.
36. Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, et al. A chromosome conformation capture ordered sequence of the barley genome. Nature. 2017;544(7651):426-+.
37. Zimin A, Puiu D, Hall R, Kingan S, Clavijo B, Salzberg S. The first near-complete assembly of the hexaploid bread wheat genome. Triticum Aestivum Gigasci. 2017;6(11):1–7.
38. Burton J, Adey A, Patwardhan R, Qiu R, Kitzman J, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 2013;31(12):1119–25.
39. Vurture G, Sedlazeck F, Nattestad M, Underwood C, Fang H, Gurtowski J, Schatz M. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics (Oxford, England). 2017;33(14):2202–4.
40. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2010;26(5):589–95.
41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics (Oxford, England). 2009;25(16):2078–9.
42. McGinnis S, Madden T. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004;32:W20-25.
43. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics (Oxford, England). 2011;27(6):764–70.
44. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol. 2018;36(12):1174–82.
45. Nurk S, Walenz B, Rhie A, Vollger M, Logsdon G, Grothe R, Miga K, Eichler E, Phillippy A, Koren S. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 2020;30(9):1291–305.
46. Roach M, Schmidt S, Borneman A. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19(1):460.
47. Guan D, McCarthy S, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics (Oxford, England). 2020;36(9):2896–8.
48. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics (Oxford, England). 2018;34(18):3094–100.
49. Kim D, Langmead B, Salzberg S. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
50. Ramírez F, Bhardwaj V, Arrigoni L, Lam K, Grüning B, Villaveces J, Habermann B, Akhtar A, Manke T. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun. 2018;9(1):189.
51. Majoros W, Pertea M, Salzberg S. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics (Oxford, England). 2004;20(16):2878–9.
52. Ter-Hovhannisyan V, Lomsadze A, Chernoff Y, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 2008;18(12):1979–90.
53. Han Y, Wessler S. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. 2010;38(22): e199.
54. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35(Web Server issue):W265-268.
55. Ou SJ, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2018;176(2):1410–22.
56. Chan P, Lin B, Mak A, Lowe T. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic acids research. 2021;49(16):9077–96.
57. Lowe T, Eddy S. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):955–64.
58. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li C, Wei L. KOBAS 20: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 2011;39:316–22.
59. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy S, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. 2005;33:D121-124.
60. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 2004;32:W309-312.
61. Han B, Jing Y, Dai J, Zheng T, Gu F, Zhao Q, Zhu F, Song X, Deng H, Wei P, et al. A chromosome-level genome assembly of Dendrobium Huoshanense using long reads and hi-C data. Genome Biol Evol. 2020;12(12):2486–90.
62. Delcher A, Salzberg S, Phillippy A. Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinform. 2003;Chapter 10:Unit 10.13.
63. Tsanakas G, Manioudaki M, Economou A, Kalaitzis P. De novo transcriptome analysis of petal senescence in Gardenia jasminoides Ellis. BMC Genomics. 2014;15(1):554.
64. Li L, Stoeckert C, Roos D. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89.
65. Chen B, Silvestri G, Dahne J, Lee K, Carpenter M. The cost-effectiveness of nicotine replacement therapy sampling in primary care: a Markov cohort simulation model. J Gen Intern Med. 2022;37(14):3684–91.
66. Nakamura T, Yamada K, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics (Oxford, England). 2018;34(14):2490–2.
67. Höhler D, Pfeiffer W, Ioannidis V, Stockinger H, Stamatakis A. RAxML Grove: an empirical phylogenetic tree database. Bioinformatics (Oxford, England). 2022;38(6):1741–2.
68. Kozlov A, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics (Oxford, England). 2019;35(21):4453–5.
69. Blanc G, Wolfe K. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell. 2004;16(7):1667–78.
70. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones S, Marra M. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–45.

71. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 1977;267(5608):275–6.
72. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 2005;22(12):2472–9.
73. Grimholt U. Whole genome duplications have provided teleosts with many roads to peptide loaded MHC class I molecules. BMC Evol Biol. 2018;18(1):25.
74. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–91.
75. Wang YP, Tang HB, DeBarry JD, Tan X, Li JP, Wang XY, Lee TH, Jin HZ, Marler B, Guo H, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40(7):14.
76. Lynch M, Conery J. The evolutionary fate and consequences of duplicate genes. Science (New York, NY). 2000;290(5494):1151–5.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 12 August 2022
doi: 10.3389/fgene.2022.844385

A High-Quality Genome Assembly of *Sorghum dochna*

Yu Chen[1,2], Yongbai Zhang[1,2], Hongjie Wang[1,2], Juan Sun[1,2], Lichao Ma[1,2], Fuhong Miao[1,2], Zixin Zhang[2], Yang Cheng[3], Jianwei Huang[4], Guofeng Yang[1,2] and Zengyu Wang[1,2]*

[1]College of Grassland Science, Qingdao Agricultural University, Qingdao, China, [2]Key Laboratory of National Forestry and Grassland Administration on Grassland Resources and Ecology in the Yellow River Delta, Qingdao Agricultural University, Qingdao, China, [3]College of Animal Science, Qingdao Agricultural University, Qingdao, China, [4]Berry Genomics Corporation, Beijing, China

Sweet sorghum (*Sorghum dochna*) is a high-quality bio-energy crop that also serves as food for humans and animals. However, there is little information on the genomic characteristics of *S. dochna*. In this study, we presented a high-quality assembly of *S. dochna* with PacBio long reads, Illumina short reads, high-throughput chromosome capture technology (Hi-C) sequencing data, gene annotation, and a comparative genome analysis. The results showed that the genome of *S. dochna* was assembled to 777 Mb with a contig N50 of 553.47 kb and a scaffold N50 of 727.11 kb. In addition, the gene annotation predicted 37,971 genes and 39,937 transcripts in the genome of *S. dochna*. A Venn analysis revealed a set of 7,988 common gene annotations by integrating five databases. A Cafe software analysis showed that 191 gene families were significantly expanded, while 3,794 were significantly contracted in *S. dochna*. A GO enrichment analysis showed that the expanded gene families were primarily clustered in the metabolic process, DNA reconstruction, and DNA binding among others. The high-quality genome map constructed in this study provides a biological basis for the future analysis of the biological characteristics of *S. dochna*, which is crucial for its breeding.

Keywords: *Sorghum dochna*, genome, assembly, comparative genome analysis, Hi-C

OPEN ACCESS

Edited by:
Luis Herrera-Estrella,
Texas Tech University, United States

Reviewed by:
Victor A Albert,
University at Buffalo, United States
Chi Yang,
Fujian Academy of Agricultural Sciences, China

****Correspondence:***
Guofeng Yang
yanggf@qau.edu.cn

Specialty section:
This article was submitted to Plant Genomics, a section of the journal Frontiers in Genetics

Received: *28 December 2021*
Accepted: *24 May 2022*
Published: *12 August 2022*

Citation:
Chen Y, Zhang Y, Wang H, Sun J, Ma L, Miao F, Zhang Z, Cheng Y, Huang J, Yang G and Wang Z (2022) A High-Quality Genome Assembly of Sorghum dochna. Front. Genet. 13:844385. doi: 10.3389/fgene.2022.844385

INTRODUCTION

Sorghum dochna belongs to the Gramineae family and has high sugar content in its stalks. Typically, it is a perennial crop except in frost-prone areas. Relevant historical records indicate that *S. dochna* was initially grown in India and Myanmar. During the mid-19th century, the United States introduced the *S. dochna* variety "Amber" from south China for cultivation, resulting in the annual production of *S. dochna* syrup as high as 111.56 million liters. Currently, *S. dochna* is cultivated in all continents of the world (Gnansounou et al., 2005).

Considering that most global economies are moving toward low-carbon energy sources, bio-renewable energy may replace oil and coal (Antonopoulou et al., 2008). *S. dochna* is an ideal bio-energy crop owing to its high photosynthetic efficiency, high resistance to stress, high sugar content (Erdei et al., 2009), high yield, and drought resistance. Thus, *S. dochna* can be used as an excellent silage material. In addition, it tastes delicious and is suitable for livestock consumption. Moreover, *S. dochna* as human food can be eaten raw or used as a raw material for making sugar, wine, and other related products. After threshing, *S. dochna* tassels can also be used to make brooms and cookware. Currently, *S. dochna* is economically valuable. There is a need for more insight into its biological mechanisms and genomic characteristics to more efficiently utilize its biological value.

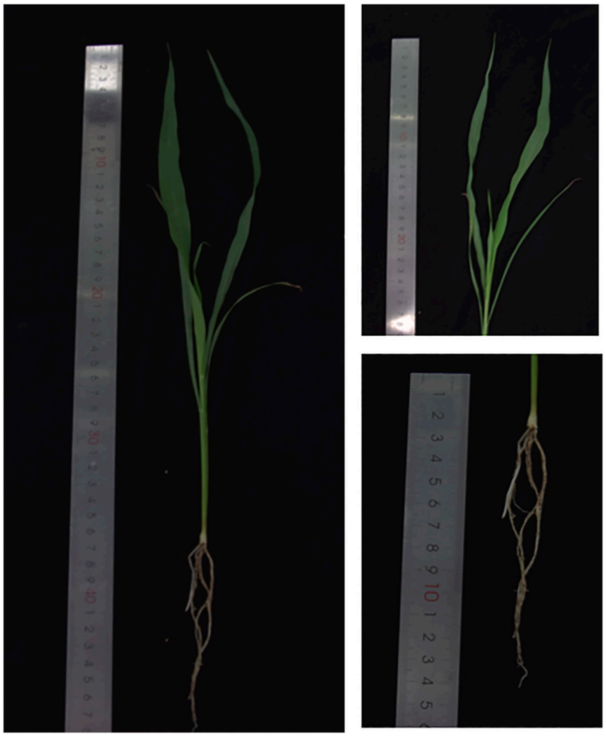


FIGURE 1 | Morphological characteristics of *Sorghum dochna* as shown in photographs that display a whole plant, leaf, and root.

To date, many studies have provided sufficient data for elucidating the *S. bicolor* genome (Paterson et al., 2009). In addition, the exploitation of *S. bicolor* as human food has increased worldwide. However, studies on the genomic analysis of *S. dochna* are limited. Moreover, the genomic characteristics of *S. dochna* are poorly understood. In this study, a genome assembly of *S. dochna* was constructed at the chromosome level using PacBio long-read, Illumina short-read, and high-throughput chromosome capture technology (Hi-C) sequencing data (Jin et al., 2021). This study provides valuable genomic data that can be used to conduct further research on the economic value of *S. dochna*. In addition, the findings of this study will facilitate comparative genomic analyses with other Gramineae forage plants.

MATERIALS AND METHODS

Materials Collection

S. dochna variety De Sheng was selected and cultivated in soil at the Research Center of Grassland, Agriculture, and Animal Husbandry of Qingdao Agricultural University (Qingdao, China). The *S. dochna* seeds were washed once with distilled water and disinfected with 75% alcohol for 1 min and NaClO for 7–8 min. After that, the seeds were dried and planted in sterilized nutrient soil. The leaves of 45-day-old seedlings were harvested (**Figure 1**), frozen in liquid nitrogen, and then stored at −80°C for subsequent analysis.

DNA and RNA Extraction

Total genomic DNA was extracted from the leaves using a Tiangen DNAsecure Novel Plant Genomic DNA Extraction Kit (Dp320-03) according to the manufacturer's instructions (Tiangen, Beijing, China). Total RNA was extracted using an EASYspin Plus Polysaccharide Polyphenols/Complex Plant RNA Rapid Extraction Kit following the manufacturer's instructions.

Survey Analysis

Raw sequence data generated by the Illumina platform (San Diego, CA, United States) were filtered by the following criteria: filtered reads with adapter sequences, filtered reads with N bases >3, and filtered reads with low-quality bases (≤5) more than 20% (Li et al., 2009). The K-mer analysis was performed using jellyfish to estimate the genome size and sample heterozygosity. The genome size can be estimated using the K-mer analysis (Chikhi and Medvedev, 2014). The distribution of K-mer depends on the characteristic of the genome and follows Poisson distribution. We estimated the genome size of *S. dochna* using the following formula: genome size = (total number of 17-mer)/(position of peak depth).

Genome Assembly and Quality Validation

Hi-Fi Assembly

We constructed a PCR-free SMRTbell library by repairing and connecting the high-quality purified genome and sequencing it by PacBio (Menlo Park, CA, United States) SMRT technology. After the library was constructed, its size was detected using an Agilent 2100 (Agilent Technologies, Santa Clara, CA, United States) fragment analyzer capillary electrophoresis or pulsed field electrophoresis. After the library was calculated by a PacBio calculator, sequencing primers and sequencing enzymes were combined into the SMRTbell template in proportion and then sequenced by diffusion loading. To obtain high-fidelity reads (Hi-Fi reads), we used SMRTlink software to conduct the subreads obtained previously for circular consensus sequencing (CCS) processing. The main parameters were min passes = 3 and min RQ = 0.99 (Sim et al., 2022).

The original data after sequencing were filtered and then assembled with hifiasm (Feng et al., 2021). First, an all vs. all comparison was used to correct the sequencing error. Second, after correction, a read overlap comparison was used again to construct a phased string graph. Finally, the contigs were generated according to the overlapping graph. The final genome sequence was obtained after de heterozygosity to generate de pseudo contigs (Koren et al., 2018).

Hi-C Assisted Genome Assembly

Raw image data files sequenced by a high-throughput sequencer (Illumina HiSeq 2500) were analyzed by base calling and transformed into sequenced reads. Raw sequencing data were stored in the FASTQ (Fq) file format. The raw reads obtained by sequencing contained a small number of articulated, repetitive, and low-quality reads, which could have affected the quality of comparison and the subsequent analysis. Therefore, we filtered the raw data to obtain clean reads. A total of 10,000 pairs of sequenced reads were randomly selected from the Hi-C

TABLE 1 | Genome assembly results.

Parameter	Contig	Scaffold
Genome assembly and Hi-C results	144	82
Total number	777,990,620	778,026,804
Total length (bp)	55,347,497	43,657,906
N50 length (bp)	43.90	43.90
GC (%)	11,660,912	—
Contig N90 length (bp)	—	72,771,365
Scaffold N50 length (bp)	—	94.09
Chromosome length (%)	—	—

Hi-C, high-throughput chromosome capture technology.

sequencing library data and compared to the NT database using BLAST. The top 10 matched species in the output results were sorted and outputted to check for bacterial contamination. JUICER software was used to compare the Hi-C data with the sketched genome. Finally, the results of the Hi-C library were compared and analyzed using 3D DNA software (Durand et al., 2016). The scaffold number was obtained using these methods.

Genome Annotation

For repeat element annotations, software RepeatMasker was used to mask the predicted repeats and known repeats (RepBase) in the genome. We used MITE Hunter, LTRharvest, LTR Finder, LTR retriever, and RepeatModeler to predict repeat sequences (Scott and Madden, 2004; Xiong et al., 2017).

We used reference protein sequences and RNA-Seq analysis to predict gene models. *Ab initio* gene prediction and annotation were performed by Augustus v3.318, SNAP, and GlimmerHMM. Augustus V3.0.3 combined with RNA-Seq data was used to predict the gene structure. First, parameters were trained with the training set. Intron hints were then obtained based on the comparison between RNA-Seq reads and the Scaffold (TopHat V2.0.10) (i.e., predicted intron location information) and then combined with intron hints for gene structure prediction. Second, SNAP and GlimmerHMM were used to predict the gene structure. The parameters were first trained with the training set, and then the genetic structure of the Scaffold shielded with repeated sequences was predicted (Ian, 2004). Third, Genemark-ET V4.57 combined with intron hints obtained from Augustus V3.0.3 was used to predict the genomic structure of the scaffold with repetitive sequences. The published protein sequences of *Oryza sativa*, *Zea mays*, *Echinochloa crus-galli*, *Brachypodium distachyon*, *S. bicolor*, and *Puccinellia tenuiflora* (NCBI) were used to perform homologous searches by GeMoMa-1.6.1.

For non-coding RNA prediction, we used tRNAscan-SE to predict the tRNA. rRNA and other types of ncRNA were searched with the Rfam database, and the specific information of ncRNA was obtained through similarity comparison.

For the gene functional annotation of protein-coding genes, we used six databases, including NR, Swiss-Prot, eggNOG, GO, KEGG, and InterPro, to perform function prediction. All these predictions of functions were integrated.

In this study, the corresponding gene function annotation results were obtained by comparing and analyzing a single database. Finally, a Venn analysis was performed by integrating the five databases to obtain the precise gene function annotation information.

Comparative Genomic Analysis

Colinear Analysis and Phylogenetic Tree

MUMmer software can be used to quickly compare two genome sequences (Delcher et al., 2003). MUMmer was used to conduct genomic colinearity analysis on *S. dochna* and its related species *S. bicolor*. The parameter was "NucMER-G 1000-C90-L200."

To identify the gene protein family, the OrthoMCL cluster analysis was adopted (Li et al., 2003). We performed all-VS-all BLAST alignments on protein-coding sequences of all the selected species (e-value was 1e^{-5} by default), calculated the similarity between sequences, and conducted a cluster analysis using the Markov clustering algorithm with an expansion coefficient of 1.5. The results of the protein family clustering were obtained. A Venn diagram was used to display the clustering results, which distinguished the endemic/common genes. The time standard point (correction point) was from the Timetree website.

Single-copy genes of each species were selected as reference markers for species with incomplete evolutionary studies, and quadruple degenerate sites were chosen to construct hypergenes. MAFFT software was used for multiple sequence comparisons of the hypergenes, and the most suitable base substitution model was selected. A phylogenetic tree was constructed based on the maximum likelihood method (ML) using RAxML software. Based on the single-copy gene family, McMctree (Burn-in = 5,000,000, sample-number = 1,000,000, and sample-frequency = 50) was used to estimate the differentiation time. The time standard point (correction point) was from the Timetree website (Hahn et al., 2005).

Gene Family Contraction and Expansion Analysis

Cafe software was used to analyze the gene families. This software can capture the changes in gene families between species based on random survival and death models combined with statistical inference methods. The number of contractions and expansions of gene families on each branch of evolution was obtained. We also determined whether contractions and expansions occurred in each gene family (Hahn et al., 2005; Hahn et al., 2007).

Positive Selected Gene Family Analysis

Early studies used the method of two-sequence alignment on all codons and the whole time zone before the divergence of the two sequences. The average value was utilized to calculate Ka and Ks. However, in fact, the vast majority of codons of a functional protein are in the process of evolution, and they are conservative most of the time. If a positive selection occurs, it will only affect some bits, and the positive selection only occurs during a specific time period. In 2002, a new method called the branch site model analysis method was reported, which can detect the positive events that occur in a specific evolutionary branch and affect only some positive selections (Jianzhi et al.). We used this method to detect the positive selection in protein coding sequences.

First, one-to-one orthology proteins from research species and related species were selected. Second, homologous protein

TABLE 2 | Statistics of the results of a comparison of the DNA library.

Sample name	Reads number	Mapped	Properly paired Mapped	Mapped DifferentChr	Mapped different ChrMapQ>=5	Secondary reads
Sorghum dochna	294,034,737	292,796,203 99.58%	279,197,414 95.55%	9,547,568 3.2%	4,598,204 1.6%	1,848,927 0.6%

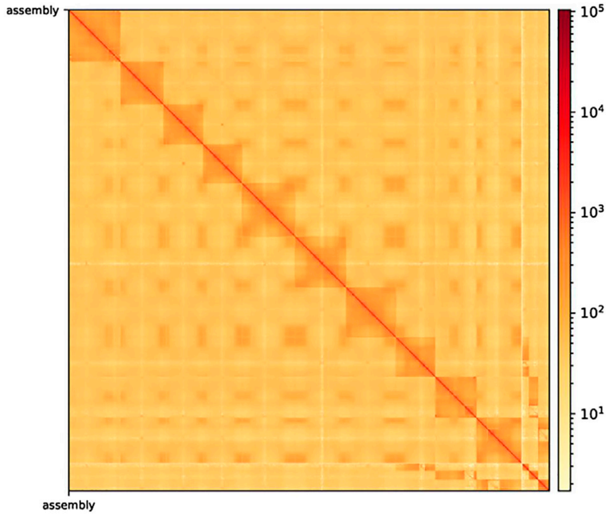


FIGURE 2 | Hi-C-assisted assembly genome interaction heatmap. The exchange within chromosomes is stronger than that between chromosomes. Moreover, the exchange between the same chromosomes with a close physical location is stronger than the exchange between chromosomes with a distant physical location. Hi-C, high-throughput chromosome capture technique.

sequences were compared with PRANK using the default parameters. Third, alignment results were filtered with G blocks with the following parameters: -t = c-e = . ft-b4 = 5-d = y. Fourth, CODEML in PAML was used to test the positive selection in a specific branch, which only affected some loci. Fifth, the Chi2 program in PAML was used to check and correct multiple hypotheses. Main parameters include degree of freedom = 2.

Based on these methods, we obtained the positive selection genes and proceeded with the GO enrichment analysis.

Whole-Genome Duplication

Whole-genome duplication (WGD) is typically associated with the rapid loss of repeated fragments, chromosome rearrangement, and the process of rearrangement back to the diploid. In this study, the distribution of synonymous substitutions (Ks) of each synonymous locus between adjacent homologous genes in the genome was constructed to detect WGD. We used BLASTP to compare the longest protein sequence of the gene in *S. dochna* genome and MCScanX to filter the comparison results. In addition, we used the yn00 tool in the PAML software package to calculate the synonymous replacement rate. The density distribution with the value of Ks was plotted for all paralog gene pairs. This approach is also known as the duplicate age distribution method (Vanneste et al., 2013). Synonymous mutations are generally considered neutral and gradually accumulate in the genome at a nearly constant rate. Therefore, Ks can represent collateral homologous genes.

RESULTS

High-Quality Gene Assembly

The quality control results of the offline data revealed 43.7 Gb of clean bases with a GC content of 43.52% and 146,092,905 clean reads (**Supplementary Table S9**). A K-mer analysis revealed that *S. dochna* is a heterozygous species (0.619%), and the 17-mer frequency distribution plot is shown as **Supplementary Figure S1**.

PacBio long reads (Nurk et al., 2020) and Illumina short reads (Dudchenko et al., 2017; Vurture et al., 2017) technologies were used to assemble the *S. dochna* genome. The PacBio clean subread statistical results are shown in **Supplementary Table S7**. We assembled the genome sequences into 1,628 contigs with a total length of 831.09 Mb, a contig N50 length of 533.94 kb, and the longest contig of 822.40 kb after the initial assembly (**Supplementary Table S1**). Thus, contigs with an average GC content of 55–57% (abnormal GC content peak in the figure) were processed by filtering the contigs. We obtained the final genome as 144 contigs with a total length of 777.99 Mb, a contig N50 of 553.47 kb, and the longest contig of 822.40 kb, which is 47 Mb bigger than that of *S. bicolor* (~730 Mb), suggesting a close relationship between *S. dochna* and *S. bicolor*. The results of the Hi-C library were analyzed using 3D DNA software, and the results revealed a genome that was 778.03 Mb long with scaffold N50 of 727.11 kb. The third-generation assembly results are shown in **Table 1**. The evaluation results of the Benchmarking Universal Single-Copy Ortholog (BUSCO) analysis indicated 97.5% completeness. A complete Single-Copy BUSCO further validated the high degree of completeness of the *S. dochna* genome assembly (**Supplementary Table S2**).

The length distribution statistics of Hi-Fi reads and Hi-Fi read bases are shown in **Supplementary Figures S4A,B**. The length of most Hi-Fi reads was distributed between 1,000 and 2,000. The gene assembly results demonstrate the high quality of the *S. dochna* genome assembly.

The completeness and accuracy of the assembly quality were assessed using the sequence data return ratio, GC-depth evaluation, and BUSCO evaluation. First of all, the results of our second-generation return ratio showed a mapping ratio of 99.58%, suggesting that most of the *S. dochna* genome had been assembled (**Table 2**). Second, an evaluation of the depth of GC found that there were no separate scattered clusters on the figure,

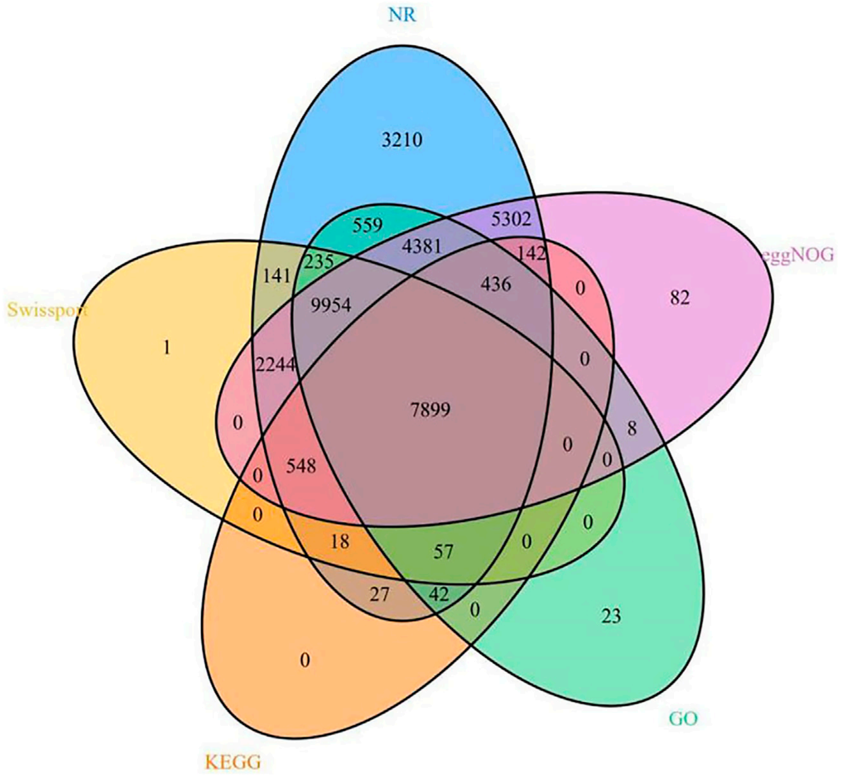


FIGURE 3 | A Venn analysis of gene function annotation.

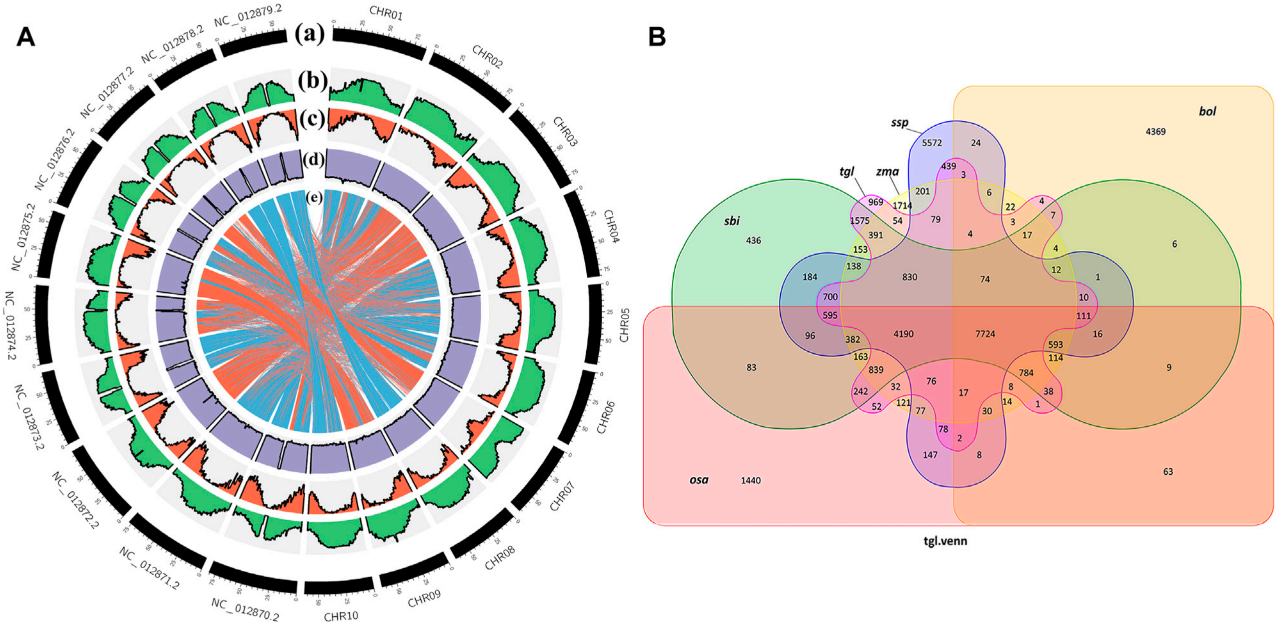


FIGURE 4 | (A) Circos display of the important features of the assembled *Sorghum dochna* genome. From outside to inside, **(A)** chromosome, **(B)** repeat sequence distribution, **(C)** gene distribution, **(D)** GC content distribution, and **(E)** colinearity between *S. dochna* and *S. bicolor*. **(B)** Venn diagram of the protein families. tgl: *S. dochna* (*S. bicolor dochna*), sbi: *S. bicolor* (*S. bicolor bicolor*), osa: rice (*Oryza sativa*), zma: maize (*Zea mays*), ssp: sugarcane (*Saccharum spontaneum*), and bol: kale (*Brassica oleracea*).

which proved that our assembly results were not polluted. A BUSCO evaluation was used to evaluate the completeness of the *S. dochna* genome (Waterhouse et al., 2018).

After Hi-C assembly, 10 chromosomes were assembled, and 751 Mb genomes were fixed to further verify the accuracy of the assembly results. This included 94.09% gene content and involved calculating the exchange between and within chromosomes. The heatmap in **Figure 2** shows the intergenomic exchange (Zhang et al., 2013).

Genome Annotation

Repeat Sequence Statistics

The results of annotation showed that Class I retrotransposons accounted for the highest proportion of the repeated sequences. The long terminal repeats (LTRs) were the most abundant transposable elements (TEs). LTR-retrotransposons accounted for 59.22%, and Gypsy accounted for 47.46% in the LTR–retrotransposons. In contrast, Copia accounted for 6.79%. Notably, Gypsy-type and Copia-type TEs accounted for most of the LTRs (**Supplementary Table S6**). Non-LTR-retrotransposons accounted for 5.97%, whereas Class II DNA transposons accounted for 9.3% of the repeated sequences (**Supplementary Table S4**).

Coding Gene and Non-Coding RNA Predictions

We predicted that 37,971 genes were encoded, and there were 39,937 transcripts in *S. dochna*. In addition, the number of genes in *S. dochna* was higher than the number of genes in *O. sativa*, *Z. mays*, *E. crus-galli*, *B. distachyon*, *S. bicolor*, and *P. tenuiflora* (**Supplementary Table S3**), which was similar to that annotated for *S. bicolor*, indicating that its genome is more complex. Based on the open reading frames, we predicted 20,108 genes in *S. dochna* (**Supplementary Table S5**). Moreover, according to the types of ncRNAs, the results of the ncRNA classes are shown in **Supplementary Table S6**. There were three sRNAs, 3,101 rRNAs, 172 miRNAs, and 847 tRNAs. There were 5,694 snRNA:: snoRNA:: CD-Box in the ncRNA.

Gene Functional Annotation

We annotated 35,309 types of gene information using six databases (NR, Swiss-Prot, eggNOG, GO, KEGG, and InterPro). The corresponding gene function annotation results were obtained by comparing the analyses of a single database. A total of 35,195 types of gene information were annotated by NR (**Supplementary Table S13**), and 21,097 types of gene information were annotated by Swiss-Prot (**Supplementary Table S14**). A total of 9,169 types of gene information were annotated by KEGG (**Supplementary Table S12**), and 23,594 types of gene information were annotated by GO (**Supplementary Table S11**). A total of 30,996 types of gene information were annotated by eggNOG (**Supplementary Table S10**). Finally, a Venn analysis was conducted by integrating the five databases (NR, Swiss-Prot, eggNOG, GO, and KEGG), which revealed a set of 7,988 common gene annotations (**Figure 2** and **Supplementary Table S15**). Venn analysis of gene functional annotations was shown in **Figure 3**.

Genome Comparison

Colinearity and Phylogenetic Relationships

When species are closely related, there is greater coverage of colinear segments on the genome, and the colinear relationship between the genomes of different species is more accurate (Krzywinski et al., 2009). **Figure 4A** shows that the colinear relationship between *S. dochna* and *S. bicolor* is relatively strong, and their relationship is relatively close. Circos displays the important features of the assembled *S. dochna* genome.

Gene Protein Family Identification and Positive Selection Gene Analysis

A protein family is a group of proteins with certain similarities in sequence and function. A protein family clustering analysis (of predicted gene proteins) groups proteins with identical or similar functions together, thus reducing the complexity of further analyses. The comparison with exogenous organisms also helps to understand and predict the gene functions. In the current genome of the Gramineae members *S. bicolor*, *S. officinarum*, *Z. mays*, and *O. sativa*, which are closely related to *S. dochna*, since they have substantial continuity in genome assembly continuity, they are selected for the protein family analysis. Among them, *S. officinarum* and *S. dochna* have the same biological characteristics of high sugar content. Simultaneously, distant species *Brassica oleracea* was selected for comparison. **Figure 4B** shows the Venn diagram of protein clustering in *S. dochna* and other species. Among them, 969 gene families were specific to *S. dochna*. A total of 1,440 were specific to *O. sativa*, 1,714 were specific to *Z. mays*, 436 were specific to *S. bicolor*, 5,572 were specific to *S. officinarum*, and 4,369 were specific to *B. oleracea*. Distributions of the numbers of single-copy genes, multi-copy genes, endemic genes, and other types of genes per species are shown in **Supplementary Figure S2**.

Phylogenetic trees that were constructed based on protein clustering results showed that *S. dochna* was closer to *S. bicolor*, while it was the most distant from *B. oleracea* (**Figure 5**). Based on the differentiation time of species, *S. bicolor* and *S. dochna* diverged from sugarcane (*Saccharum officinarum*) 9.2 and 1.4 million years ago, respectively. A cafe software analysis showed that 191 gene families were significantly expanded, while 3,794 were significantly contracted in *S. dochna* (tgl) after the family-wide *p*-value threshold was 0.05. The result of GO enrichment in expanded gene families is shown in **Supplementary Figure S3**. A GO function enrichment analysis of these gene families revealed that the expanded gene families were primarily clustered in the metabolic process, DNA reconstruction, and DNA binding among others (**Supplementary Figure S5**). The positive selection analysis model with *S. dochna* as the foreground branch and other species as the background branch was established. Finally, we obtained four significant positive selected genes. The GO enrichment analysis showed that these positive selection genes were primarily clustered in organic cyclic compound binding, nucleic acid binding, nucleotidyl transferase activity, and tRNA methylation among others (**Supplementary Figures S5, S6**). One significant positive selection gene was clustered in peptidyl-prolyl *cis-trans* isomerase (PPIase)

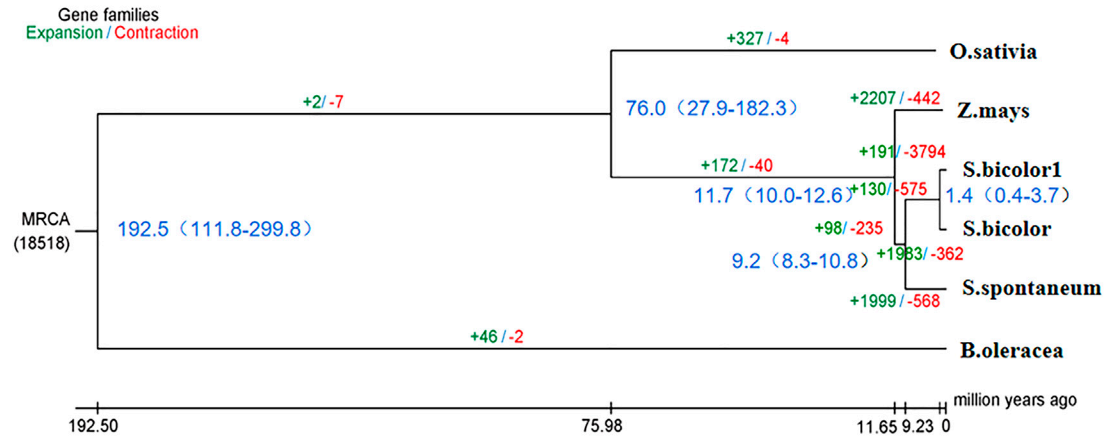


FIGURE 5 | Phylogenetic tree of the species. In the analysis that estimated the time of differentiation of species, the branch length obtained is the base replacement rate, and after the analysis of species differentiation time, the branch length is the time in million years. *O. sativa*: *Oryza sativa*. *Z. mays*: *Zea mays*. *S. bicolor1*: *Sorghum dochna*. *S. bicolor*: *Sorghum bicolor. S. spontaneum*: *Saccharum spontaneum. B. oleracea*: *Brassica oleracea.*

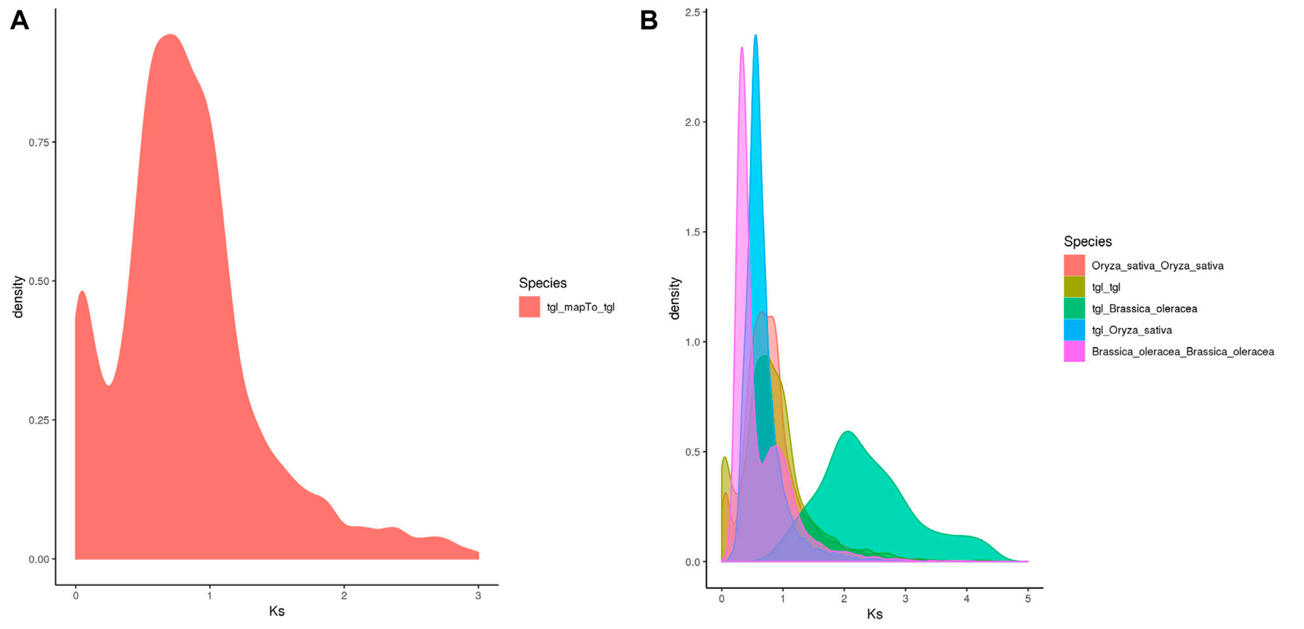


FIGURE 6 | (A) Ks distribution map of *Oryza sativa*. The peaks indicate that WGD events occurred during the evolution of species. **(B)** Ks distribution map of the complete genome replication of *Sorghum dochna* and its related species. Tgl: *S. dochna*. WGD, whole-genome duplication.

activity. The GO enrichment information is shown in **Supplementary Table S8**.

According to the species differentiation time, *S. dochna* and *S. bicolor* diverged 1.4 million years ago. During this period, the temperature of the Earth was lower by 5–10°. Since then, the Earth has undergone several alterations in climate. *S. bicolor* is native to Africa, while *S. dochna* is native to India/Myanmar, which is currently separated by the Indian Ocean (Dutt, 1999). Therefore, it is hypothesized that the formation and differentiation of the two *S. bicolor* species could be related to the climate and tectonic plate movement at that time (Chase, 1978). However, we did not explore the similarities in physiological functions and gene family clustering between the two *S. bicolor* species in more detail. Therefore, further studies should be conducted to fully elucidate their specific biological properties (Hahn et al., 2005).

During the GO enrichment analysis, one significant positive selection gene was clustered in peptidyl-prolyl *cis-trans* isomerase (PPIase) activity (**Supplementary Figure S6**). PPIase can catalyze the conformation of protein substrates or the N-terminal of proline residues in the polypeptide from a homeopathic structure to a trans structure (Maruyama et al., 2000). This type of protein can also improve the stress resistance of plants when they are in adversity and pass on the stress resistance to future generations. Therefore, it is hypothesized that the high stress resistance of *S. dochna* is related to the positive selection of this gene. Other positive selection genes were clustered in tRNA methylation. tRNA methylation primarily occurs in the nitrogen atom of tRNA and can also occur in the oxygen atom of the 2′ hydroxyl of nucleotide ribose ring (Gustilo et al., 2008). In addition, the 5′ carbon atom on purine and the 2′ and 8′ carbon atoms on adenosine have also been identified. The methylation phenomenon is primarily related to protein translation and the stability of tRNA (Motorin and Helm, 2011). In addition, for organic cyclic compound binding, *S. dochna* is a high-quality bio-energy crop with high sugar content. Most sugar structures are constituted with organic cyclic compounds, such as furan and pyran. Therefore, we hypothesized that during the evolution of sweet sorghum, positive selection genes were enriched in the binding of organic cyclic compounds, which could be used in the synthesis of sugars (Lingle et al., 2012).

Whole-Genome Duplication Analysis

Whole-genome duplication (WGD) is often associated with a rapid loss of repeated fragments and chromosome rearrangements. Notably, it provides new materials for the evolution of organisms, particularly plants, which assists them in their adaptation to new environments. A whole-genome duplication analysis performed on the pan-genome of *S. dochna* (**Figure 5**) revealed gene replication and loss and a sudden increase in the Ks within a certain period (shown as a peak), suggesting that a WGD event could have occurred. Otherwise, loss occurred (shown as a smooth decline).

The Ks of ortholog gene pairs between the *S. dochna* genome and those of related species were searched for a density distribution map. The Ks distribution of orthologs (**Figures 6A,B**) suggested that a WGD event occurred in *S. dochna* as in other species of the Gramineae family. As shown in **Figure 5A**, two differentiation events occurred in *S. dochna* when the Ks values were 0.1 and 0.8. Simultaneously, **Figure 5B** shows that the green one represents the differentiation event of *S. dochna* and *B. oleracea*, and *S. dochna* and *B. oleracea* had the highest Ks values. Therefore, the WGD event occurred the earliest in these two species, followed by *S. dochna* and *O. sativa*, while the differentiation of *O. sativa*, *B. oleracea*, and *S. dochna* occurred relatively late. Thus, the Ks value was relatively low.

CONCLUSION

In this study, we used PacBio long reads, Illumina short reads, and Hi-C sequences to assemble the *S. dochna* genome and analyze its gene family and relationship with other species. Our findings provide a preliminary understanding of the *S. dochna* genome. A high-quality chromosome assembly was achieved using PacBio long reads, Illumina short reads, and Hi-C sequences. The genome size of *S. dochna* is 777 Mb, with a contig N50 of 553.5 kb and a GC content of 43.9%. The coding gene analysis revealed 37,971 genes and 39,937 transcripts in the *S. dochna* genome.

The genome comparison indicated that *S. dochna* and *S. bicolor* had the strongest colinearity. GO enrichment revealed that the positive selection genes primarily clustered in organic cyclic compound binding, nucleic acid binding, nucleotide transferase activity, and tRNA methylation among others. However, the synthetic pathway of sugar production in *S. dochna* is still unclear (Hakim and Wijaya, 2009). Thus, subsequent studies on genome exploration should focus on the transcriptome and proteome of *S. dochna*. In addition, only one variety of *S. dochna* was used in this study. Cognizant of this, future studies should use multiple varieties to comparatively analyze the species and construct reference-quality genome sequences.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

YuC, ZW and GY conceived and designed this research. YuC analyzed data and wrote the manuscript. YuC, YZ, HW and JS executed the data analyses. JS participated in the discussion of the results. LM, HS, FM, ZZ, YaC and JH collected samples. GY, JH contributed to the evaluation and discussion of the results and manuscript revisions. All authors have read and approved the final version.

FUNDING

This study was supported by the National Nature Science Foundation of China (U1906201), Shandong Forage Research System (SDAIT-23-01), China Agriculture Research System (CARS-34), and the First Class Grassland Science Discipline Program of Shandong Province (1619002), China.

ACKNOWLEDGMENTS

The authors would like to thank GY, ZW, and JS (Professor of Grassland Science, Qingdao Agricultural University) for their help in data analysis and article writing. We are grateful for the scientific research funding provided by the College of Grassland Science of Qingdao Agricultural University and the experimental help provided by Beijing Berry and Kang.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.844385/full#supplementary-material

REFERENCES

Antonopoulou, G., Gavala, H. N., Skiadas, I. V., Angelopoulos, K., and Lyberatos, G. (2008). Biofuels Generation from Sweet Sorghum: Fermentative Hydrogen Production and Anaerobic Digestion of the Remaining Biomass. *Bioresour. Technol.* 99, 110–119. doi:10.1016/j.biortech.2006.11.048

Chase, C. G. (1978). Plate Kinematics: The Americas, East Africa, and the Rest of the World. *Earth Planet. Sci. Lett.* 37, 355–368. doi:10.1016/0012-821x(78)90051-1

Chikhi, R., and Medvedev, P. (2014). Informed and Automated K-Mer Size Selection for Genome Assembly. *Bioinformatics* 30, 31–37. doi:10.1093/bioinformatics/btt310

Delcher, A. L., Salzberg, S. L., and Phillippy, A. M. (2003). Using MUMmer to Identify Similar Regions in Large Sequence Sets. *Curr. Protoc. Bioinforma* 10. doi:10.1002/0471250953.bi1003s00

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). De Novo assembly of the *Aedes aegypti* Genome Using Hi-C Yields Chromosome-Length Scaffolds. *Science* 356, 92–95. doi:10.1126/science.aal3327

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., et al. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell. Syst.* 3, 95–98. doi:10.1016/j.cels.2016.07.002

Dutt, V. P. (1999). *India's Foreign Policy in a Changing World*. Vikas Publishing House.

Erdei, É., Pepó, P., Csapó, J., Tóth, S., and Szabó, B. (2009). Sweet Sorghum (Sorghum Dochna L.) Restorer Lines Effects on Nutritional Parameters of Stalk Juice. *Acta Agrar. debr.*, 51–56. doi:10.34101/actaagrar/36/2792

Feng, X. Cheng, H. Portik, D., and Li, H. (2021). *Metagenome Assembly of High-Fidelity Long Reads with Hifiasm-Meta*. Arxiv E-Prints.

Gnansounou, E., Dauriat, A., and Wyman, C. E. (2005). Refining Sweet Sorghum to Ethanol and Sugar: Economic Trade-Offs in the Context of North China. *Bioresour. Technol.* 96, 985–1002. doi:10.1016/j.biortech.2004.09.015

Gustilo, E. M., Vendeix, F. A., and Agris, P. F. (2008). tRNA's Modifications Bring Order to Gene Expression. *Curr. Opin. Microbiol.* 11, 134–140. doi:10.1016/j.mib.2008.02.003

Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C., and Cristianini, N. (2005). Estimating the Tempo and Mode of Gene Family Evolution from Comparative Genomic Data. *Genome Res.* 15, 1153–1160. doi:10.1101/gr.3567505

Hahn, M. W., Han, M. V., and Han, S.-G. (2007). Gene Family Evolution across 12 Drosophila Genomes. *PLoS Genet.* 3, e197. doi:10.1371/journal.pgen.0030197

Hakim, L. A., and Wijaya, I. H. (2009). Production of Bioethanol from Sweet Sorghum: A Review. *Afr. J. Agric. Res.* 4, 772–780. doi:10.1021/jf9024163

Ian, K. (2004). Gene Finding in Novel Genomes. *BMC Bioinforma.* 5, 59. doi:10.1186/1471-2105-5-59

Jin, S., Bian, C., Jiang, S., Han, K., and Fu, H. (2021). A Chromosome-Level Genome Assembly of the Oriental River Prawn, Macrobrachium Nipponense. *GigaScience* 10, 1–9. doi:10.1093/gigascience/giaa160

Koren, S., Rhie, A., Walenz, B. P., Dilthey, A. T., Bickhart, D. M., Kingan, S. B., et al. (2018). Assembly of happy-resolved genomes with trio binning. *Nat. Biotechnol.* 36, 1174–1182.

Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: An Information Aesthetic for Comparative Genomics. *Genome Res.* 19, 1639–1645. doi:10.1101/gr.092759.109

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* 13, 2178–2189. doi:10.1101/gr.1224503

Lingle, S. E., Tew, T. L., Rukavina, H., and Boykin, D. L. (2012). Post-harvest Changes in Sweet Sorghum I: Brix and Sugars. *Bioenerg. Res.* 5, 158–167. doi:10.1007/s12155-011-9164-0

Maruyama, T., Furutani, M., and Furutani, M. (2000). Archaeal Peptidyl Prolyl Cis-Trans Isomerases (PPIases). *Front. Biosci.* 5, D821–D836. doi:10.2741/maruyama

Motorin, Y., and Helm, M. (2011). RNA Nucleotide Methylation. *WIREs RNA* 2, 611–631. doi:10.1002/wrna.79

Nurk, S. Walenz, B. P. Rhie, A. Vollger, M. R. and Koren, S. (2020). *HiCanu: Accurate Assembly of Segmental Duplications, Satellites, and Allelic Variants from High-Fidelity Long Reads*. Havard university.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The Sorghum Bicolor Genome and the Diversification of Grasses. *Nature* 457. doi:10.1038/nature07723

Scott, M. G., and Madden, T. L. (2004). BLAST: at the Core of a Powerful and Diverse Set of Sequence Analysis Tools. *Nucleic Acids Res.* 32, W20–W25.

Sim, S. B., Corpuz, R. L., Simmonds, T. J., and Geib, S. M. (2022). HiFiAdapterFilt, a Memory Efficient Read Processing Pipeline, Prevents Occurrence of Adapter Sequence in PacBio HiFi Reads and Their Negative Impacts on Genome Assembly. *BMC Genomics* 23, 157. doi:10.1186/s12864-022-08375-1

Vanneste, K., Van de Peer, Y., and Maere, S. (2013). Inference of Genome Duplications from Age Distributions Revisited. *Mol. Biol. Evol.* 30 (1), 177–190. doi:10.1093/molbev/mss214

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Han, F., Gurtowski, J., et al. (2017). GenomeScope: Fast Reference-free Genome Profiling from Short Reads. *Bioinformatics* 33, 2202–2204. doi:10.1093/bioinformatics/btx153

Waterhouse, R. M., Seppey, M., Simao, F. A., Manni, M., Loannidis, P., Klioutchnikov, G., et al. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* 35, 543–548. doi:10.1093/molbev/msx319

Xiong, F., Liu, J., He, L., Han, Z., Huang, Z., Tang, X., et al. (2017). Recent Advances on the Development and Utilization of Molecular Markers Based on LTR Retrotransposons and MITE Transposons from Peanut(*Arachis hypogaea* L.). *Mol. Plant Breed.* 2.

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2013). Allele-defined Genome of the Autopolyploid Sugarcane *Saccharum Spontaneum L. Nat. Genet.* 50, 1565–1573. doi:10.1038/s41588-018-0237-2

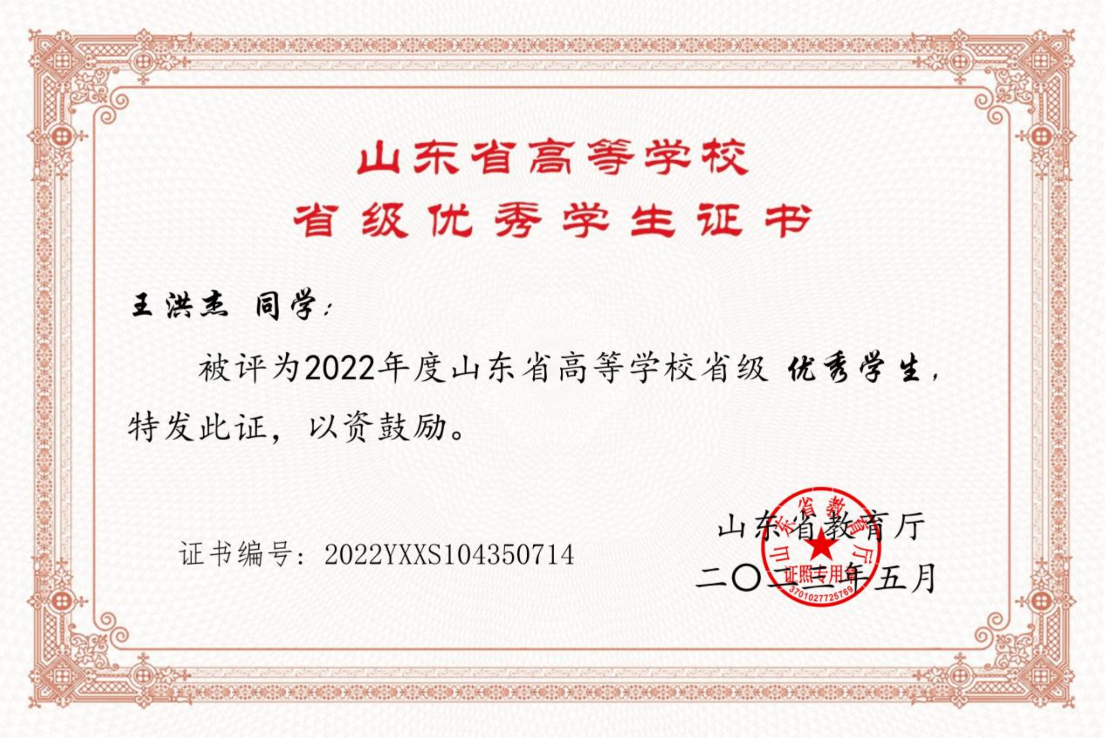
Conflict of Interest: Author JW was employed by the company Berry Genomics Corporation.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

二、省级荣誉

（一）山东省高等学校省级优秀学生（2023.05）



山东省高等学校

省级优秀学生证书

王洪杰 同学：

被评为2022年度山东省高等学校省级 **优秀学生**，特发此证，以资鼓励。

证书编号：2022YXXS104350714

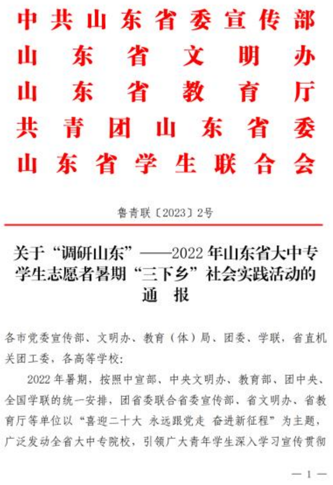
山东省教育厅

二〇二三年五月

山东省教育厅 证照专用 3701027725769

（二）山东省“三下乡”社会实践优秀学生（2023.02）

（三）山东省“三下乡”社会实践优秀服务团队（团队负责人）（2023.02）



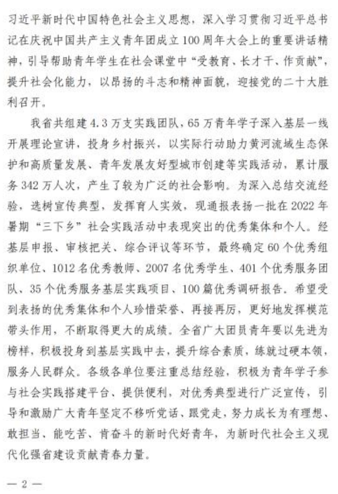
中共山东省委宣传部
山　东　省　文　明　办
山　东　省　教　育　厅
共　青　团　山　东　省　委
山　东　省　学　生　联　合　会

鲁青联〔2023〕2号

关于“调研山东”——2022年山东省大中专学生志愿者暑期“三下乡”社会实践活动的通　报

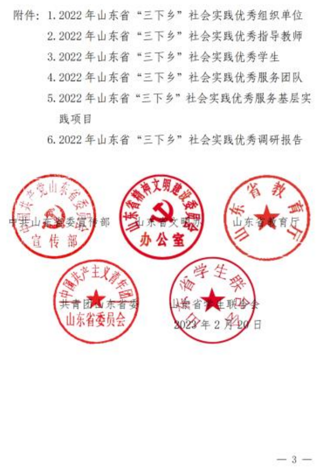
各市党委宣传部、文明办、教育（体）局、团委、学联，省直机关团工委、各高等学校：

2022年暑期，按照中宣部、中央文明办、教育部、团中央、全国学联的统一安排，团省委联合省委宣传部、省文明办、省教育厅等单位以“喜迎二十大　永远跟党走　奋进新征程”为主题，广泛发动全省大中专院校，引领广大青年学生深入学习宣传贯彻



习近平新时代中国特色社会主义思想，深入学习贯彻习近平总书记在庆祝中国共产主义青年团成立100周年大会上的重要讲话精神，引导帮助青年学生在社会课堂中“受教育、长才干、作贡献”，提升社会化能力，以昂扬的斗志和精神面貌，迎接党的二十大胜利召开。

我省共组建4.3万支实践团队，65万青年学子深入基层一线开展理论宣讲，投身乡村振兴，以实际行动助力黄河流域生态保护和高质量发展、青年发展友好型城市创建等实践活动，累计服务342万人次，产生了较为广泛的社会影响。为深入总结交流经验，选树宣传典型，发挥育人实效，现通报表扬一批在2022年暑期“三下乡”社会实践活动中表现突出的优秀集体和个人。经基层申报、审核把关、综合评议等环节，最终确定60个优秀组织单位、1012名优秀教师、2007名优秀学生、401个优秀服务团队、35个优秀服务基层实践项目、100篇优秀调研报告。希望受到表扬的优秀集体和个人珍惜荣誉、再接再厉，更好地发挥模范带头作用，不断取得更大的成绩。全省广大团员青年要以先进为榜样，积极投身到基层实践中去，提升综合素质，练就过硬本领，服务人民群众。各级各单位要注重总结经验，积极为青年学子参与社会实践搭建平台、提供便利，对优秀典型进行广泛宣传，引导和激励广大青年坚定不移听党话、跟党走，努力成长为有理想、敢担当、能吃苦、肯奋斗的新时代好青年，为新时代社会主义现代化强省建设贡献青春力量。



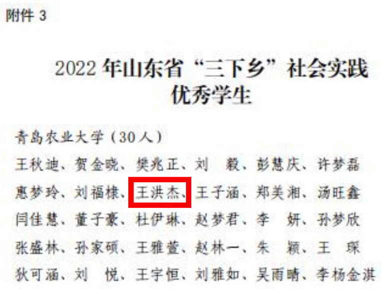
附件：1.2022年山东省“三下乡”社会实践优秀组织单位
2.2022年山东省“三下乡”社会实践优秀指导教师
3.2022年山东省“三下乡”社会实践优秀学生
4.2022年山东省“三下乡”社会实践优秀服务团队
5.2022年山东省“三下乡”社会实践优秀服务基层实践项目
6.2022年山东省“三下乡”社会实践优秀调研报告

中国共产党山东省委员会宣传部
山东省精神文明建设委员会办公室
山东省教育厅
中国共产主义青年团山东省委员会
山东省学生联合会

中共山东省委宣传部　山东省文明办　山东省教育厅

共青团山东省委　山东省学生联合会

2023年2月20日

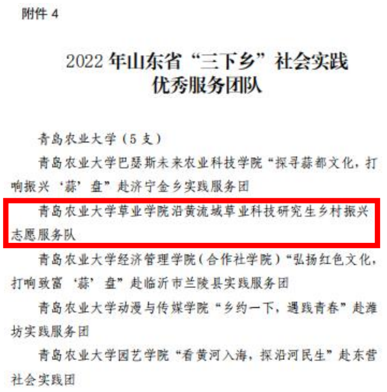


附件3

2022年山东省“三下乡”社会实践优秀学生

青岛农业大学（30人）

王秋迪、贺金晓、樊兆正、刘　毅、彭慧庆、许梦磊
惠梦玲、刘福棣、王洪杰、王子涵、郑美湘、汤旺鑫
闫佳慧、董子豪、杜伊琳、赵梦君、李　妍、孙梦欣
张盛林、孙家硕、王雅萱、赵林一、朱　颖、王　琛
狄可涵、刘　悦、王宇恒、刘雅如、吴雨晴、李杨金淇



附件4

2022年山东省“三下乡”社会实践优秀服务团队

青岛农业大学（5支）

青岛农业大学巴瑟斯未来农业科技学院“探寻蒜都文化，打响振兴‘蒜’盘”赴济宁金乡实践服务团

青岛农业大学草业学院沿黄流域草业科技研究生乡村振兴志愿服务队

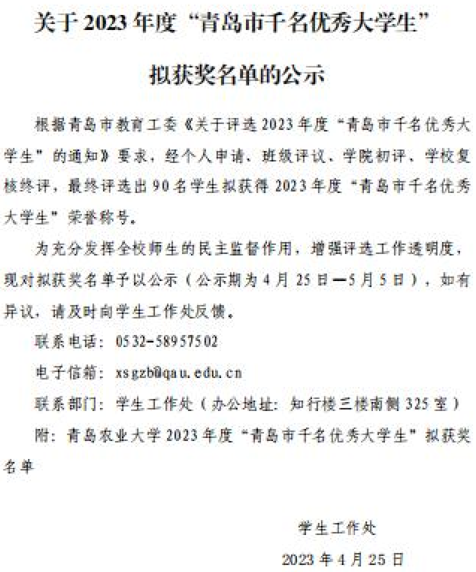
青岛农业大学经济管理学院（合作社学院）“弘扬红色文化，打响致富‘蒜’盘”赴临沂市兰陵县实践服务团

青岛农业大学动漫与传媒学院“乡约一下，遇践青春”赴潍坊实践服务团

青岛农业大学园艺学院“看黄河入海，探沿河民生”赴东营社会实践团

三、市级荣誉

（一）青岛市千名优秀大学生（2023.05）



关于2023年度“青岛市千名优秀大学生”

拟获奖名单的公示

根据青岛市教育工委《关于评选2023年度“青岛市千名优秀大学生”的通知》要求，经个人申请、班级评议、学院初评、学校复核终评，最终评选出90名学生拟获得2023年度“青岛市千名优秀大学生”荣誉称号。

为充分发挥全校师生的民主监督作用，增强评选工作透明度，现对拟获奖名单予以公示（公示期为4月25日—5月5日），如有异议，请及时向学生工作处反馈。

联系电话：0532-58957502

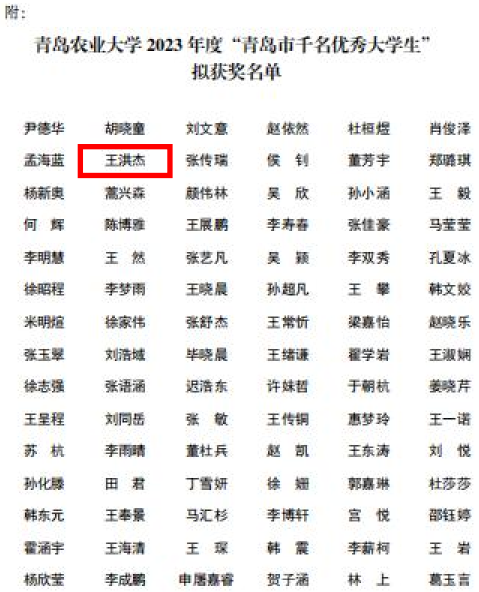
电子信箱：xsgzb@qau.edu.cn

联系部门：学生工作处（办公地址：知行楼三楼南侧325室）

附：青岛农业大学2023年度“青岛市千名优秀大学生”拟获奖名单

学生工作处

2023年4月25日



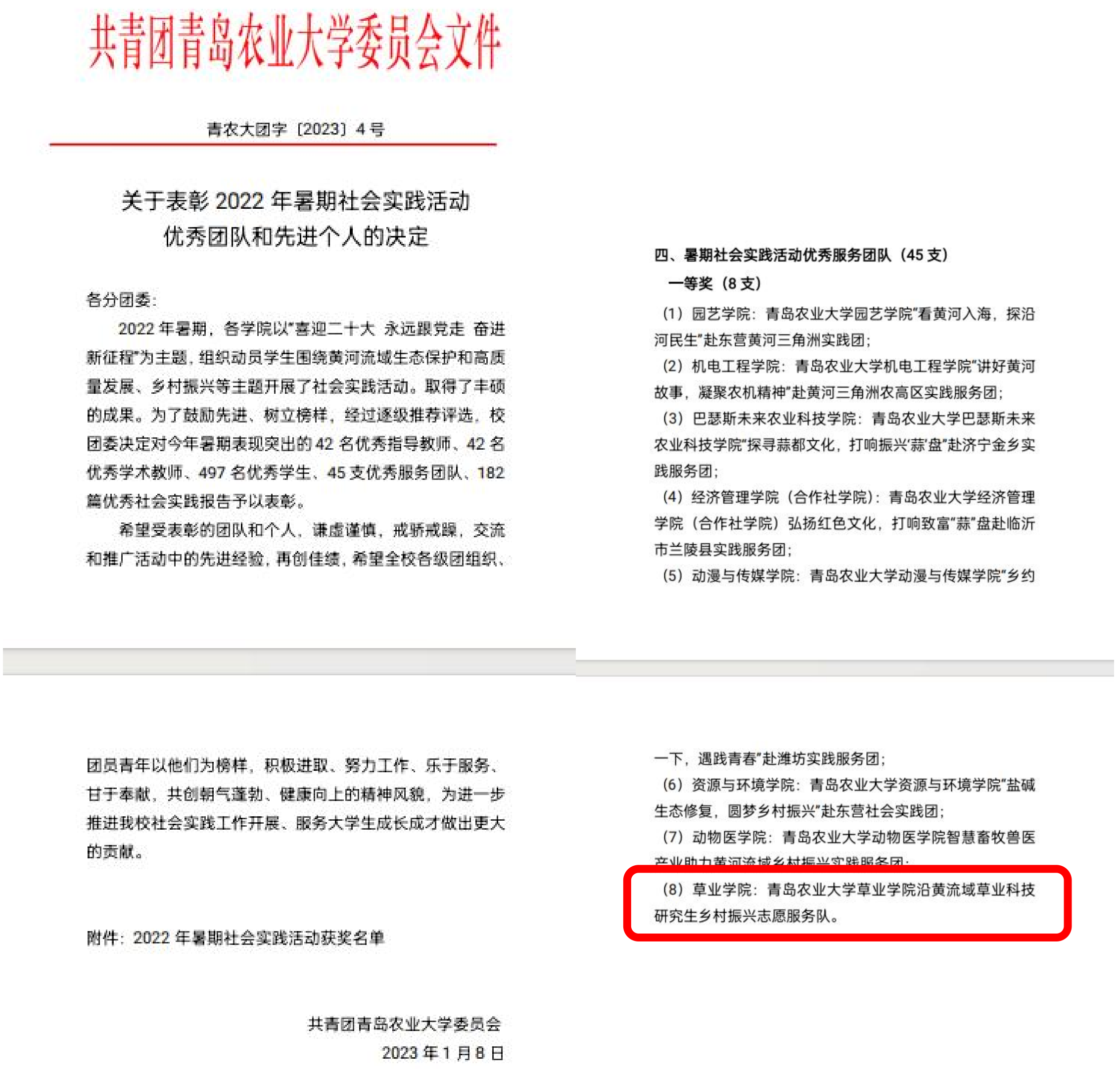
附：

青岛农业大学2023年度“青岛市千名优秀大学生”
拟获奖名单

尹德华	胡晓童	刘文意	赵依然	杜桓煜	肖俊泽
孟海蓝	王淇杰	张传瑞	侯 钊	董芳宇	郑璐琪
杨新奥	蒿兴森	颜伟林	吴 欣	孙小涵	王 毅
何 辉	陈博雅	王展鹏	李寿春	张佳豪	马莹莹
李明慧	王 然	张艺凡	吴 颖	李双秀	孔夏冰
徐昭程	李梦雨	王晓晨	孙超凡	王 攀	韩文姣
米明煊	徐家伟	张舒杰	王常忻	梁嘉怡	赵晓乐
张玉翠	刘浩城	毕晓晨	王绪谦	翟学岩	王淑娴
徐志强	张语涵	迟浩东	许姝哲	于朝杭	姜晓芹
王呈程	刘同岳	张 敏	王传钢	惠梦玲	王一诺
苏 杭	李雨晴	董杜兵	赵 凯	王东涛	刘 悦
孙化滕	田 君	丁雪妍	徐 媚	郭嘉琳	杜莎莎
韩东元	王奉景	马汇杉	李博轩	宫 悦	邵钰婷
霍涵宇	王海清	王 琛	韩 震	李蔚柯	王 岩
杨欣莹	李成鹏	申屠嘉睿	贺子涵	林 上	葛玉言

四、校级荣誉

（一）青岛农业大学2022年暑期社会实践活动优秀服务团队一等奖（团队负责人）（2023.01）



共青团青岛农业大学委员会文件

青农大团字〔2023〕4号

关于表彰2022年暑期社会实践活动
优秀团队和先进个人的决定

各分团委：

2022年暑期，各学院以"喜迎二十大 永远跟党走 奋进新征程"为主题，组织动员学生围绕黄河流域生态保护和高质量发展、乡村振兴等主题开展了社会实践活动，取得了丰硕的成果。为了鼓励先进、树立榜样，经过逐级推荐评选，校团委决定对今年暑期表现突出的42名优秀指导教师、42名优秀学术教师、497名优秀学生、45支优秀服务团队、182篇优秀社会实践报告予以表彰。

希望受表彰的团队和个人，谦虚谨慎，戒骄戒躁，交流和推广活动中的先进经验，再创佳绩，希望全校各级团组织、团员青年以他们为榜样，积极进取、努力工作、乐于服务、甘于奉献，共创朝气蓬勃、健康向上的精神风貌，为进一步推进我校社会实践工作开展、服务大学生成长成才做出更大的贡献。

附件：2022年暑期社会实践活动获奖名单

共青团青岛农业大学委员会
2023年1月8日

四、暑期社会实践活动优秀服务团队（45支）

一等奖（8支）

（1）园艺学院：青岛农业大学园艺学院"看黄河入海，探沿河民生"赴东营黄河三角洲实践团；

（2）机电工程学院：青岛农业大学机电工程学院"讲好黄河故事，凝聚农机精神"赴黄河三角洲农高区实践服务团；

（3）巴瑟斯未来农业科技学院：青岛农业大学巴瑟斯未来农业科技学院"探寻蒜都文化，打响振兴'蒜'盘"赴济宁金乡实践服务团；

（4）经济管理学院（合作社学院）：青岛农业大学经济管理学院（合作社学院）弘扬红色文化，打响致富"蒜"盘赴临沂市兰陵县实践服务团；

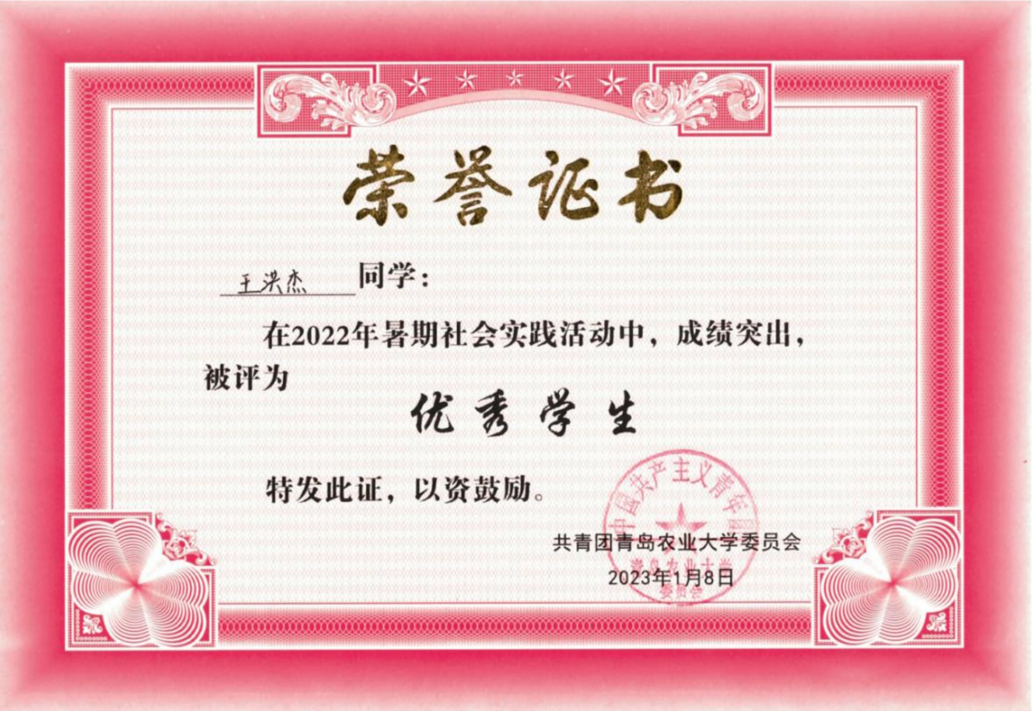
（5）动漫与传媒学院：青岛农业大学动漫与传媒学院"乡约一下，遇践青春"赴潍坊实践服务团；

（6）资源与环境学院：青岛农业大学资源与环境学院"盐碱生态修复，圆梦乡村振兴"赴东营社会实践团；

（7）动物医学院：青岛农业大学动物医学院智慧畜牧兽医产业助力黄河流域乡村振兴实践服务团；

（8）草业学院：青岛农业大学草业学院沿黄流域草业科技研究生乡村振兴志愿服务队。

（二）青岛农业大学2022年暑期社会实践活动优秀学生（2023.01）



荣誉证书

王洪杰 同学：

在2022年暑期社会实践活动中，成绩突出，被评为

优秀学生

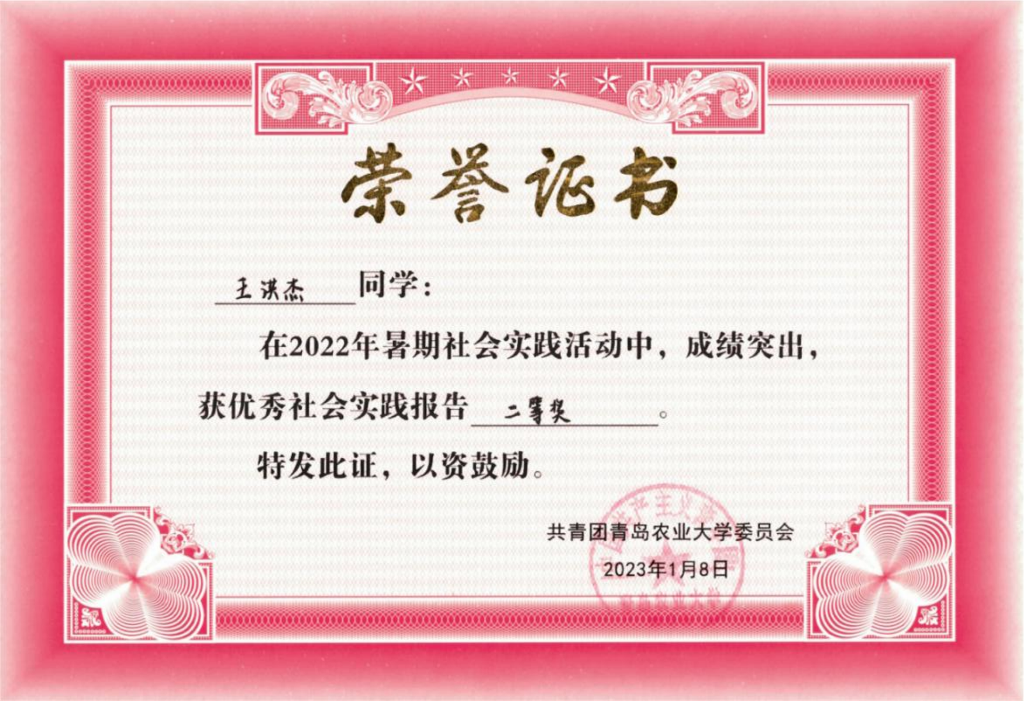
特发此证，以资鼓励。

共青团青岛农业大学委员会

2023年1月8日

中国共产主义青年团青岛农业大学委员会

（三）青岛农业大学2022年暑期社会实践活动优秀社会实践报告二等奖（2023.01）



荣誉证书

王洪杰 同学：

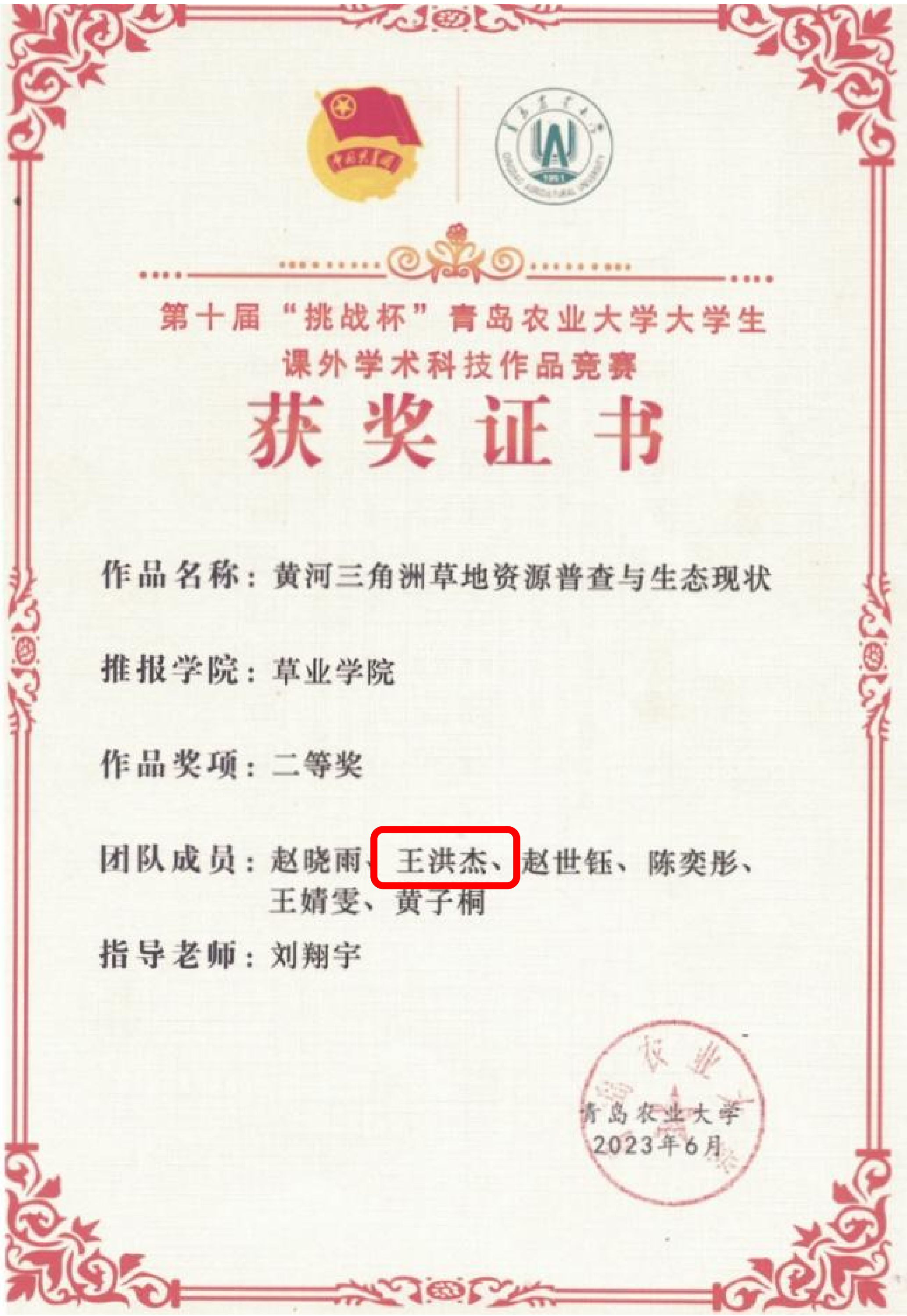
在2022年暑期社会实践活动中，成绩突出，获优秀社会实践报告 二等奖 。

特发此证，以资鼓励。

共青团青岛农业大学委员会

2023年1月8日

（四）青岛农业大学大学生课外学术科技作品竞赛二等奖（第二位）（2023.06）



第十届“挑战杯”青岛农业大学大学生课外学术科技作品竞赛

获奖证书

作品名称：黄河三角洲草地资源普查与生态现状

推报学院：草业学院

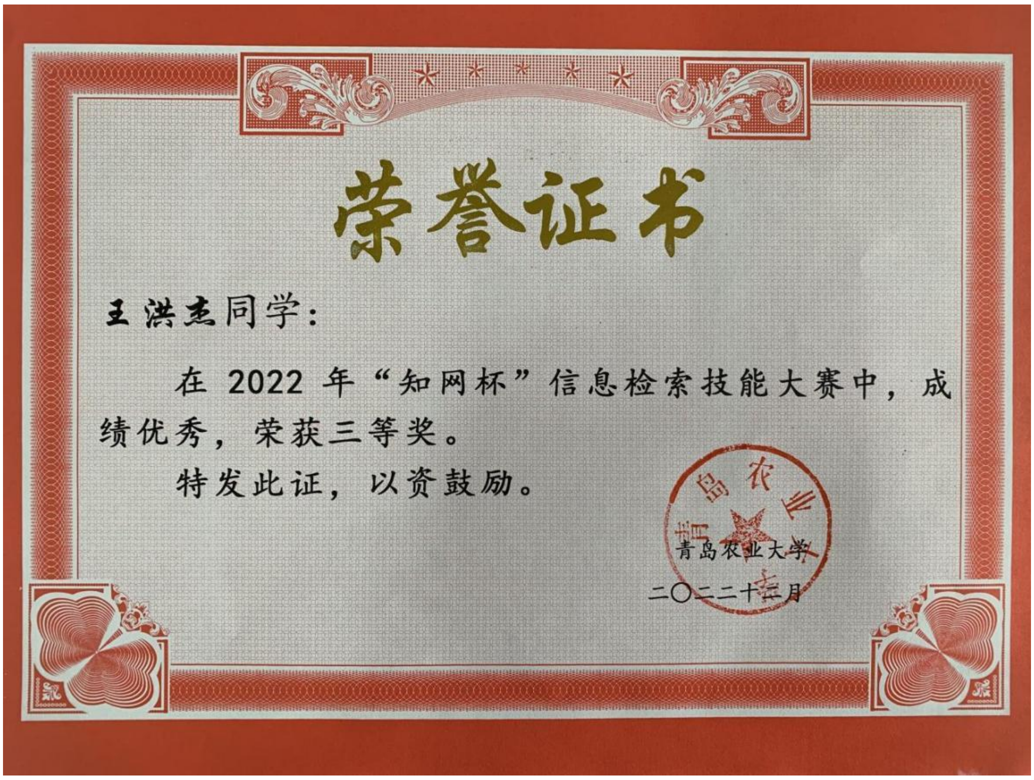
作品奖项：二等奖

团队成员：赵晓雨、王洪杰、赵世钰、陈奕彤、王婧雯、黄子桐

指导老师：刘翔宇

青岛农业大学
2023年6月

（五）青岛农业大学“知网杯”信息检索技能大赛三等奖（2022.12）



荣誉证书

王洪杰同学：

在 2022 年“知网杯”信息检索技能大赛中，成绩优秀，荣获三等奖。

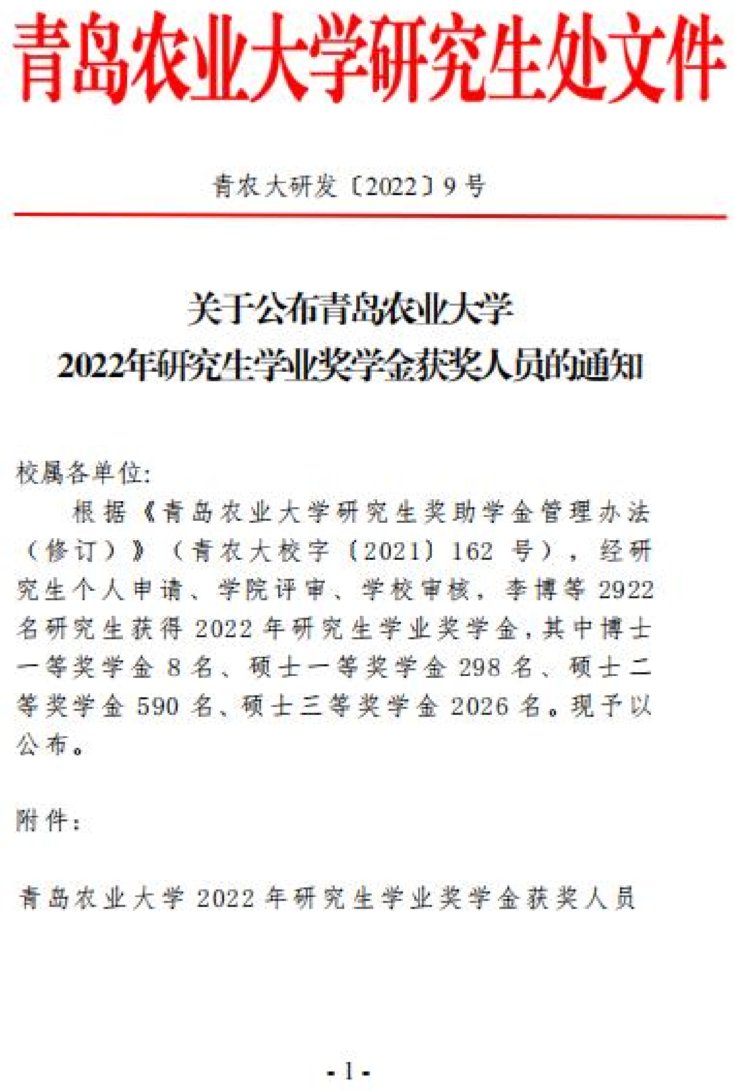
特发此证，以资鼓励。

青岛农业大学

二〇二二十二月

青岛农业大学

（六）硕士二等奖学金（2022.11）



青岛农业大学研究生处文件

青农大研发〔2022〕9号

关于公布青岛农业大学 2022年研究生学业奖学金获奖人员的通知

校属各单位:

根据《青岛农业大学研究生奖助学金管理办法（修订）》（青农大校字〔2021〕162 号），经研究生个人申请、学院评审、学校审核，李博等 2922 名研究生获得 2022 年研究生学业奖学金，其中博士一等奖学金 8 名、硕士一等奖学金 298 名、硕士二等奖学金 590 名、硕士三等奖学金 2026 名。现予以公布。

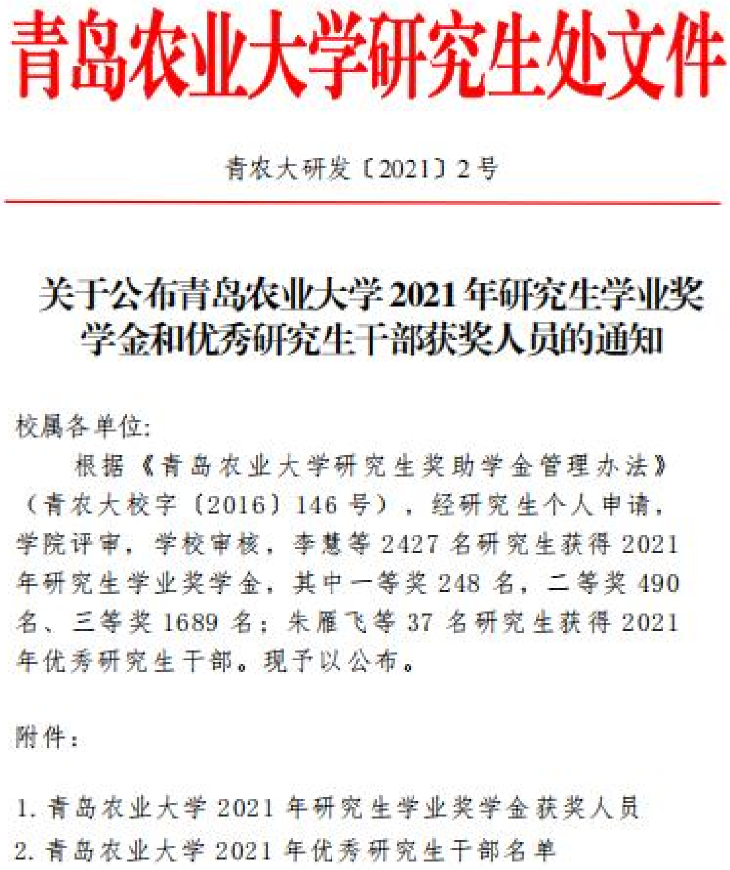
附件：

青岛农业大学 2022 年研究生学业奖学金获奖人员

- 1 -

313	王洪杰	20212203010	草学	草业学院	硕士二等奖学金

（七）硕士三等奖学金（2021.11）



青岛农业大学研究生处文件

青农大研发〔2021〕2号

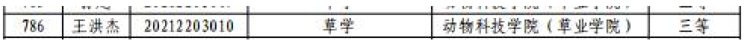
关于公布青岛农业大学2021年研究生学业奖学金和优秀研究生干部获奖人员的通知

校属各单位:

根据《青岛农业大学研究生奖助学金管理办法》（青农大校字〔2016〕146号），经研究生个人申请，学院评审，学校审核，李慧等2427名研究生获得2021年研究生学业奖学金，其中一等奖248名，二等奖490名、三等奖1689名；朱雁飞等37名研究生获得2021年优秀研究生干部。现予以公布。

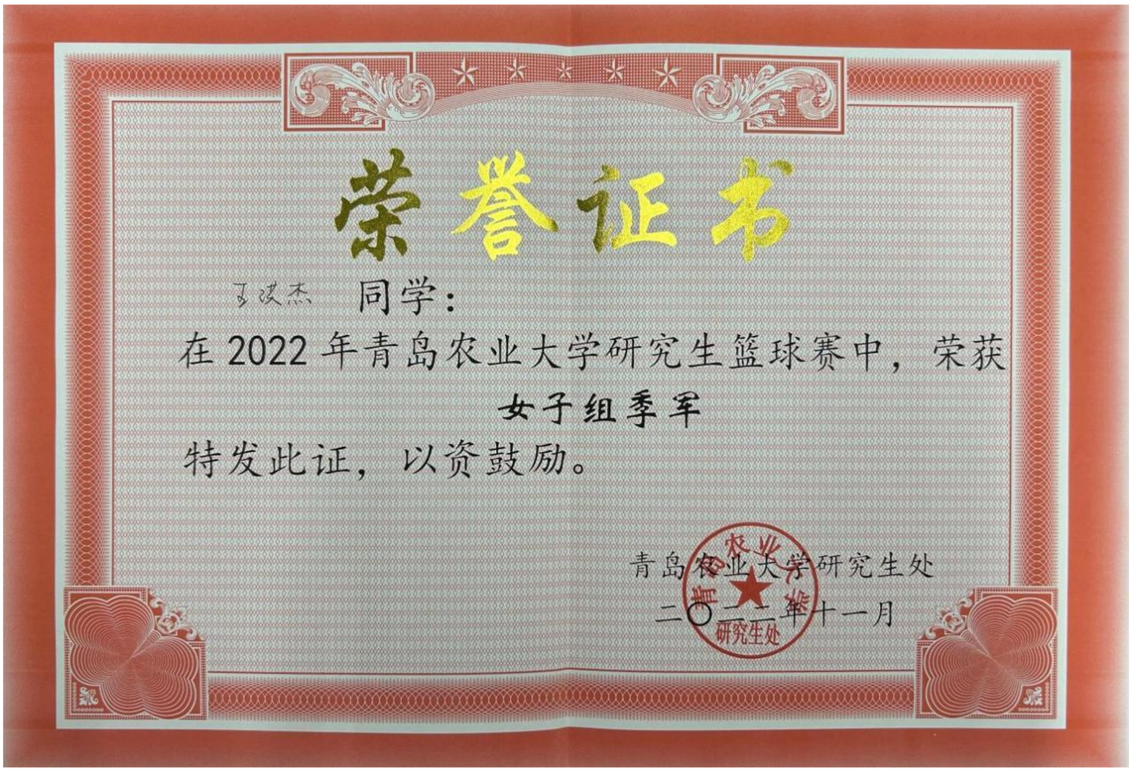
附件：

1. 青岛农业大学2021年研究生学业奖学金获奖人员
2. 青岛农业大学2021年优秀研究生干部名单



786	王洪杰	20212203010	草学	动物科技学院（草业学院）	三等

（八）青岛农业大学研究生篮球赛女子组季军（2022.11）



荣誉证书

王洪杰 同学：

在2022年青岛农业大学研究生篮球赛中，荣获

女子组季军

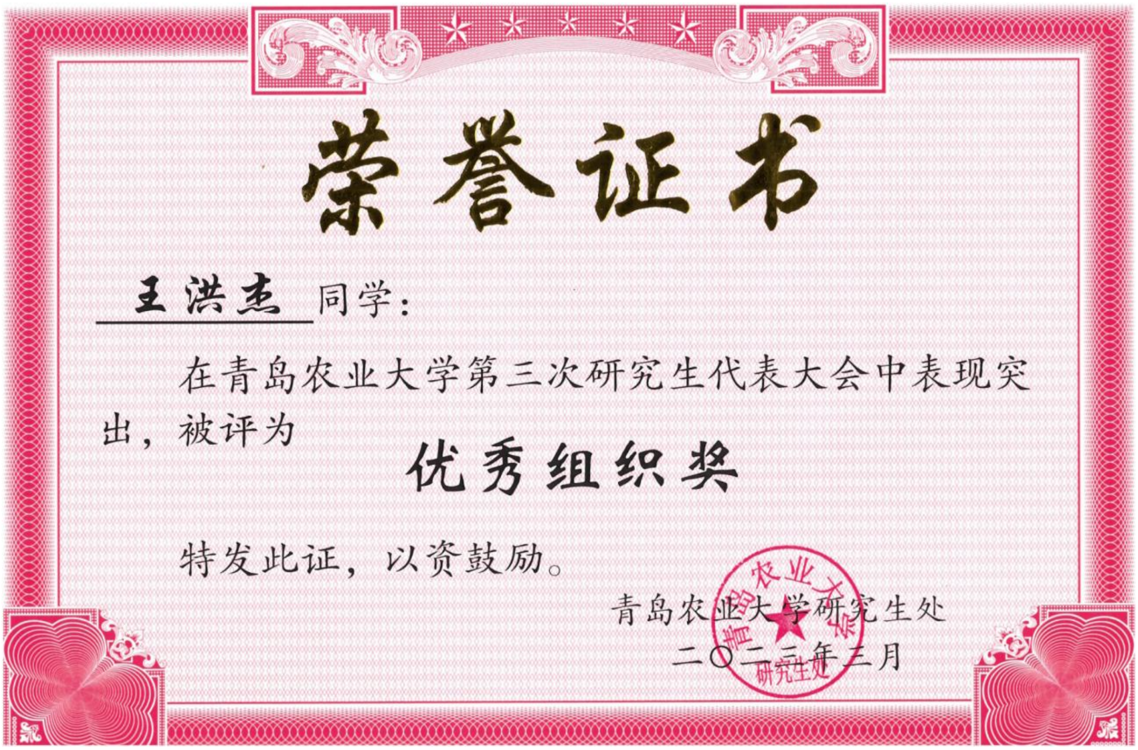
特发此证，以资鼓励。

青岛农业大学研究生处

二〇二二年十一月

青岛农业大学研究生处

（九）青岛农业大学第三次研究生代表大会优秀组织奖（2023.03）



荣誉证书

王洪杰 同学：

在青岛农业大学第三次研究生代表大会中表现突出，被评为

优秀组织奖

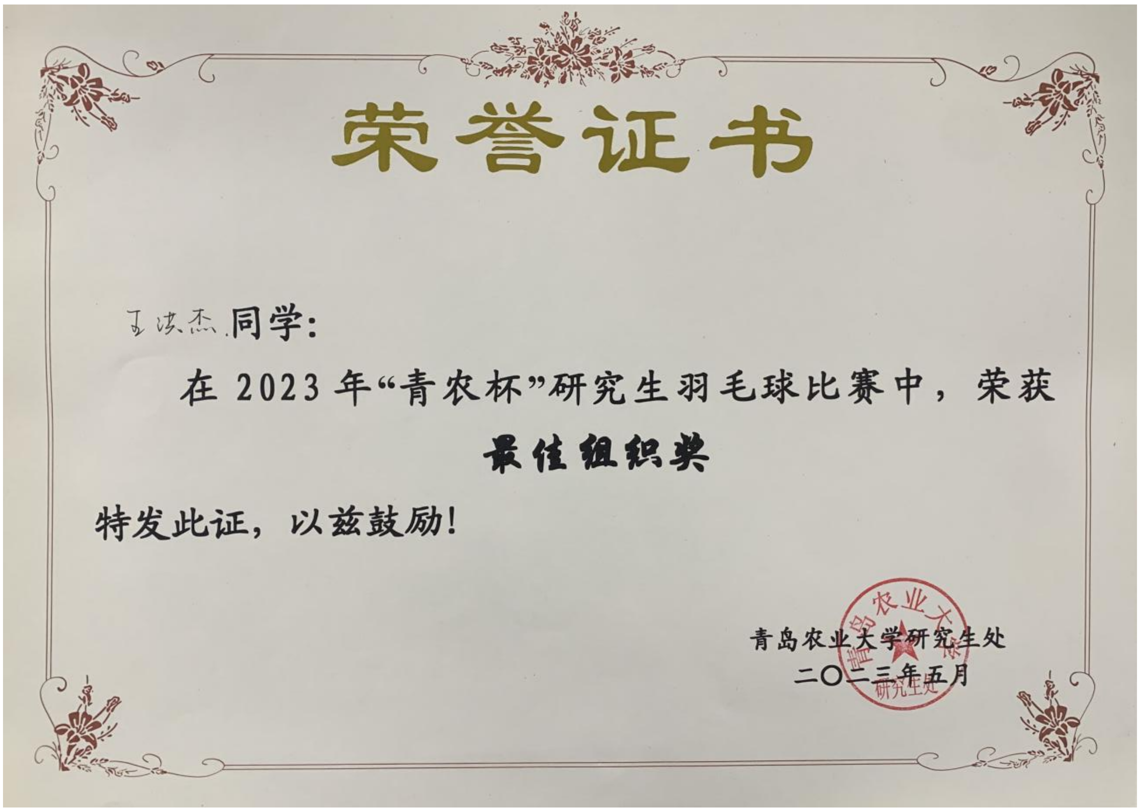
特发此证，以资鼓励。

青岛农业大学研究生处

二〇二三年三月

青岛农业大学研究生处

（十）“青农杯”研究生羽毛球比赛最佳组织奖（2023.05）



荣誉证书

王汝杰 同学：

在 2023 年“青农杯”研究生羽毛球比赛中，荣获

最佳组织奖

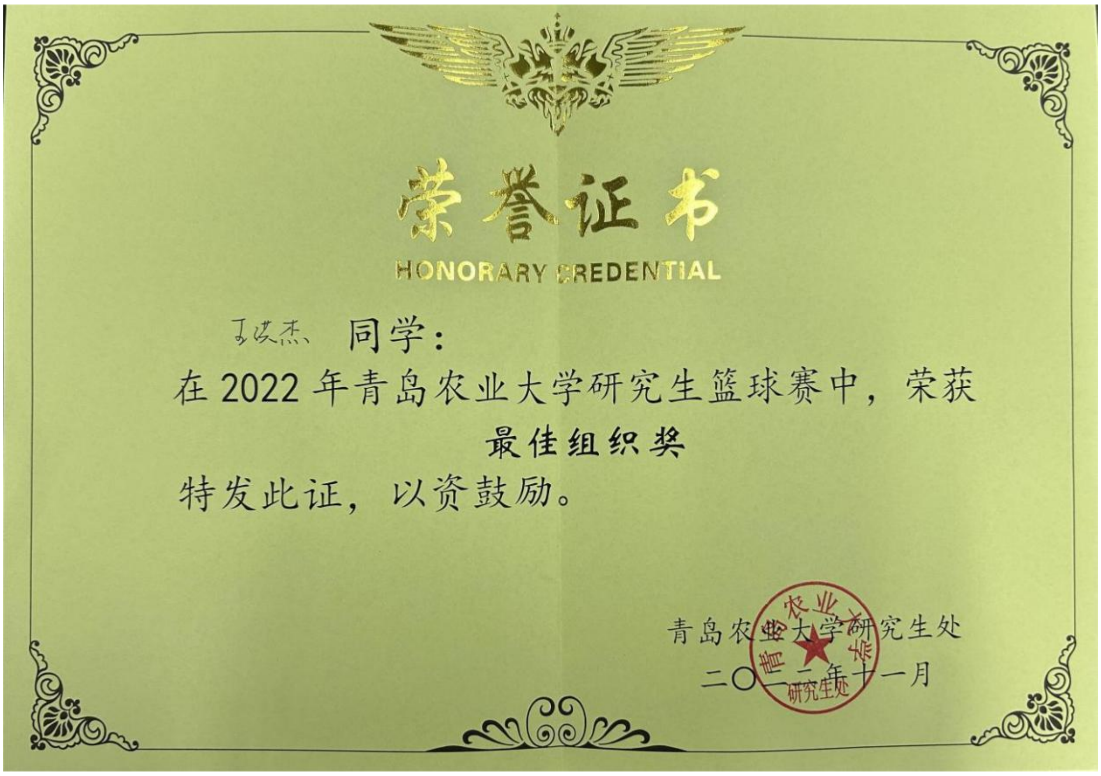
特发此证，以兹鼓励！

青岛农业大学研究生处

二〇二三年五月

青岛农业大学 研究生处

（十一）青岛农业大学研究生篮球赛最佳组织奖（2022.11）



荣誉证书

HONORARY CREDENTIAL

王汝杰 同学：

在 2022 年青岛农业大学研究生篮球赛中，荣获

最佳组织奖

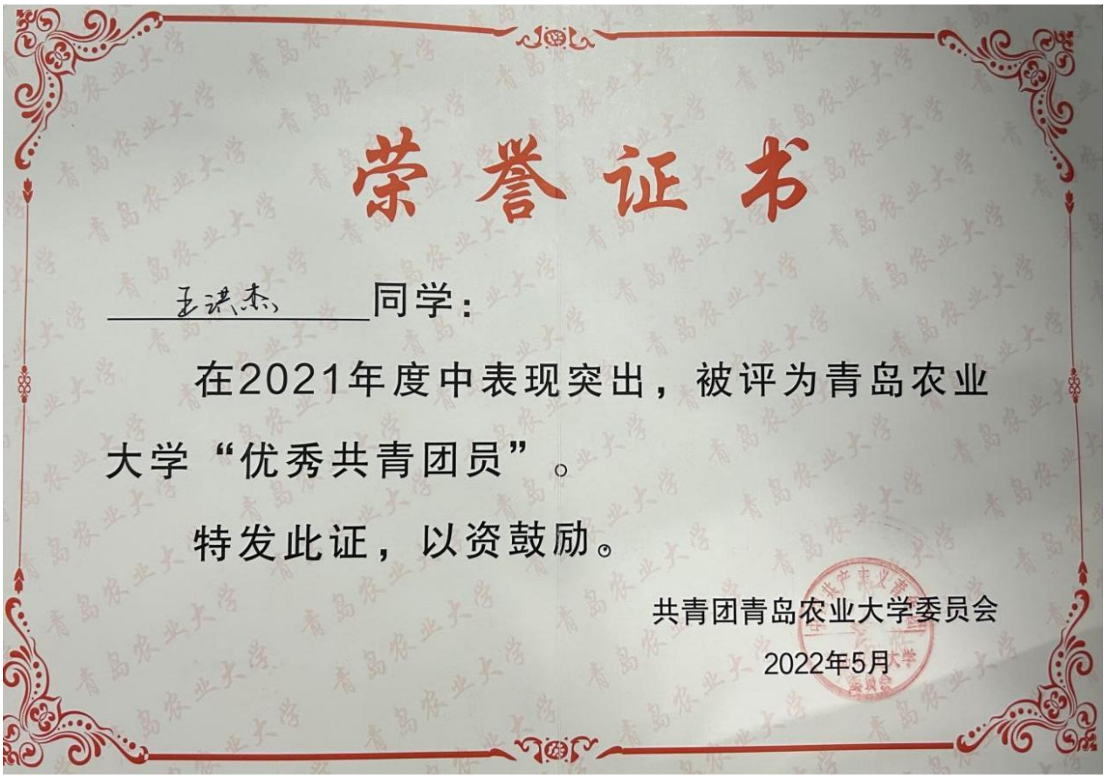
特发此证，以资鼓励。

青岛农业大学研究生处

二〇二二年十一月

青岛农业大学 研究生处

（十二）青岛农业大学“优秀共青团员”（2022.05）



荣誉证书

王洪杰同学：

在2021年度中表现突出，被评为青岛农业大学“优秀共青团员”。

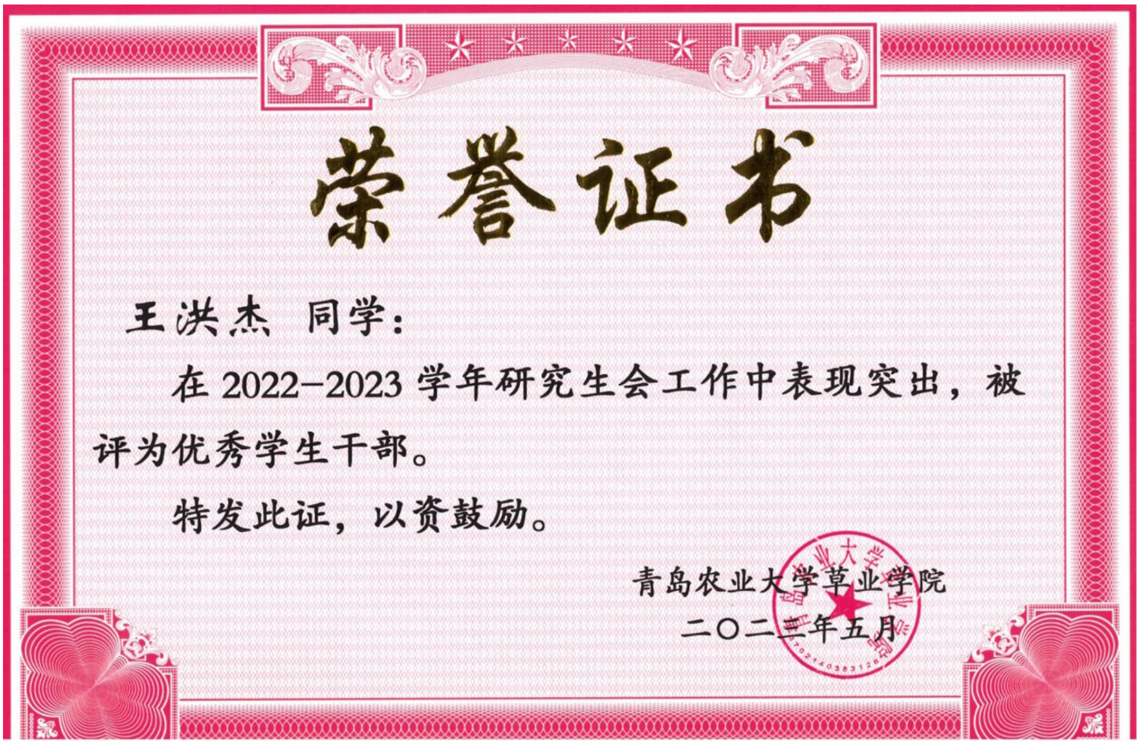
特发此证，以资鼓励。

共青团青岛农业大学委员会

2022年5月

五、院级荣誉

（一）青岛农业大学草业学院优秀学生干部（2023.05）



荣誉证书

王洪杰 同学：

在 2022-2023 学年研究生会工作中表现突出，被评为优秀学生干部。

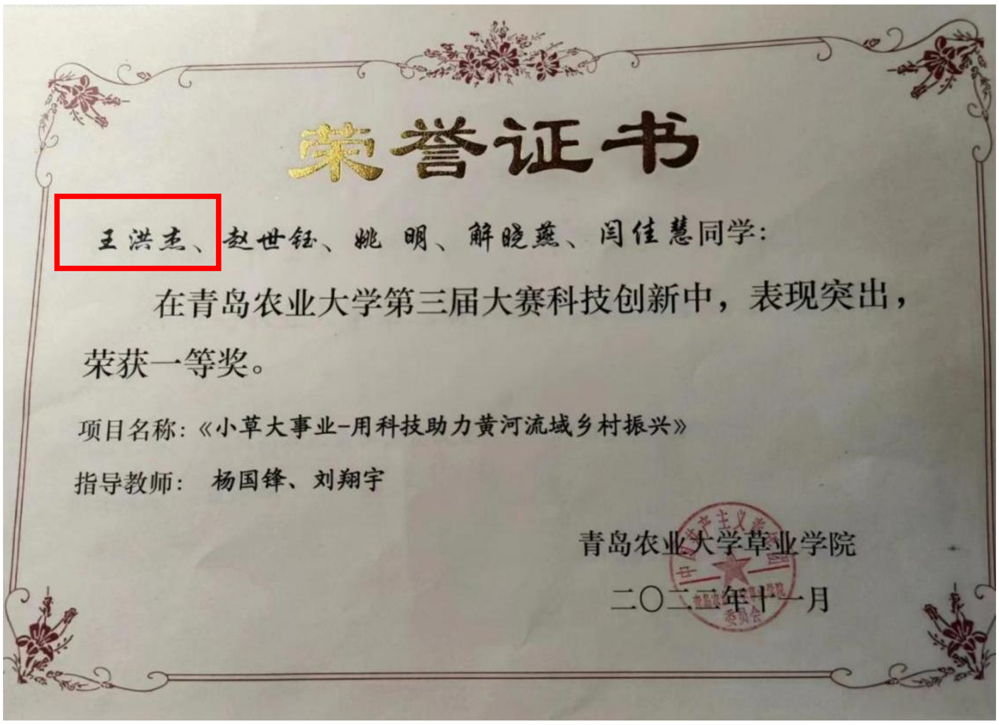
特发此证，以资鼓励。

青岛农业大学草业学院

二〇二三年五月

青岛农业大学草业学院

（二）青岛农业大学第三届大赛科技创新一等奖（第一位）（2022.11）



荣誉证书

王洪杰、赵世钰、姚　明、解晓燕、闫佳慧同学：

在青岛农业大学第三届大赛科技创新中，表现突出，荣获一等奖。

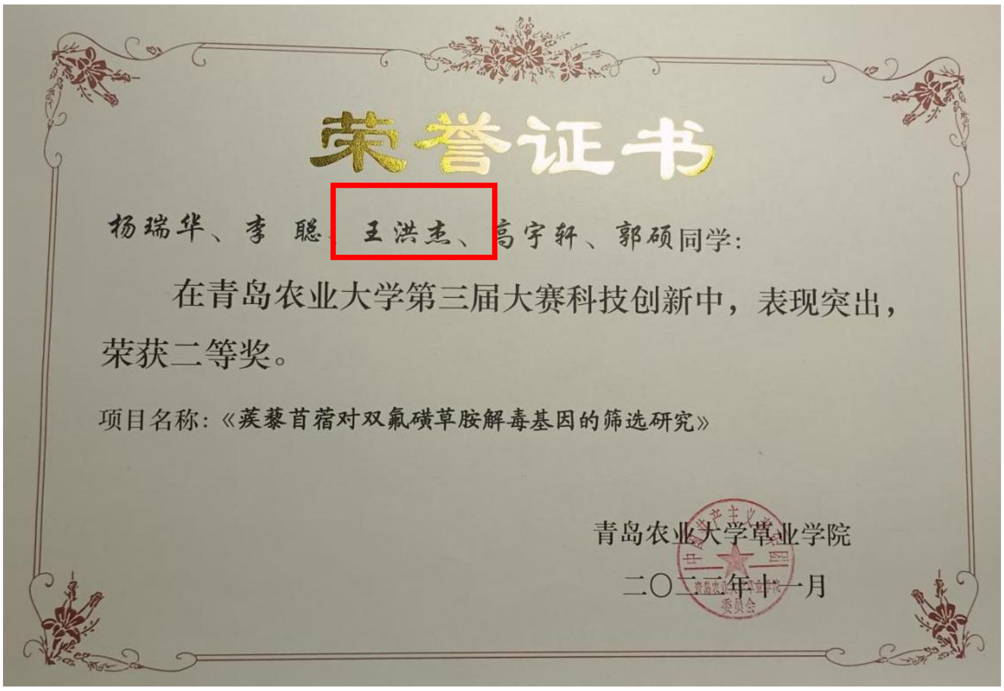
项目名称：《小草大事业-用科技助力黄河流域乡村振兴》

指导教师：　杨国锋、刘翔宇

青岛农业大学草业学院

二〇二二年十一月

（三）青岛农业大学第三届大赛科技创新二等奖（第三位）（2022.11）



荣誉证书

杨瑞华、李　聪、王洪杰、高宇轩、郭硕同学：

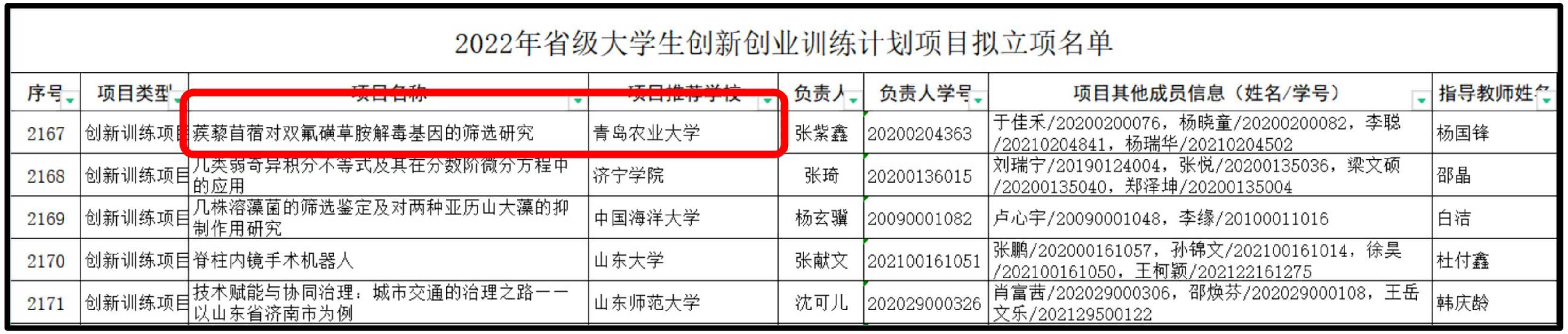
在青岛农业大学第三届大赛科技创新中，表现突出，荣获二等奖。

项目名称：《蒺藜苜蓿对双氟磺草胺解毒基因的筛选研究》

青岛农业大学草业学院

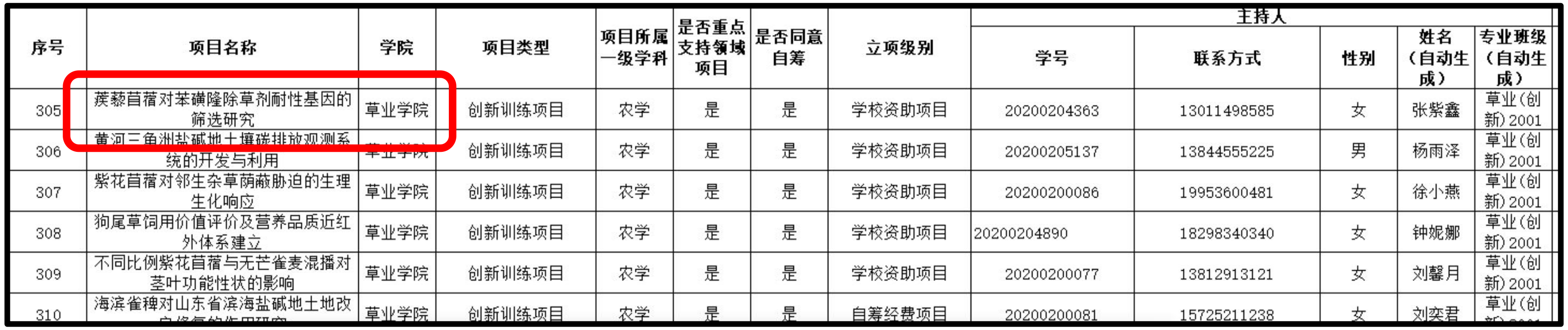
二〇二二年十一月

六、协助本科生开展科研工作（省级立项 1 项、校级立项 1 项）



2022年省级大学生创新创业训练计划项目拟立项名单

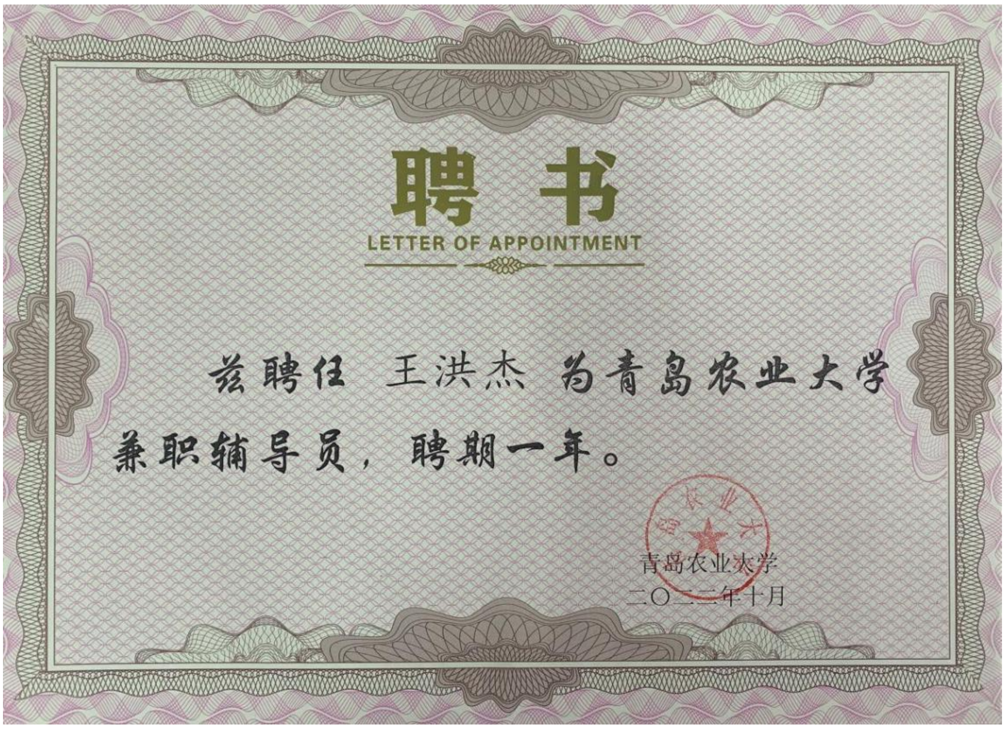
序号	项目类型	项目名称	项目推荐学校	负责人	负责人学号	项目其他成员信息（姓名/学号）	指导教师姓名
2167	创新训练项目	蒺藜苜蓿对双氟磺草胺解毒基因的筛选研究	青岛农业大学	张紫鑫	20200204363	于佳禾/20200200076，杨晓童/20200200082，李聪/20210204841，杨瑞华/20210204502	杨国锋
2168	创新训练项目	几类弱奇异积分不等式及其在分数阶微分方程中的应用	济宁学院	张琦	20200136015	刘瑞宁/20190124004，张悦/20200135036，梁文硕/20200135040，郑泽坤/20200135004	邵晶
2169	创新训练项目	几株溶藻菌的筛选鉴定及对两种亚历山大藻的抑制作用研究	中国海洋大学	杨玄骥	20090001082	卢心宇/20090001048，李缘/20100011016	白洁
2170	创新训练项目	脊柱内镜手术机器人	山东大学	张献文	202100161051	张鹏/202000161057，孙锦文/202100161014，徐昊/202100161050，王柯颖/202122161275	杜付鑫
2171	创新训练项目	技术赋能与协同治理：城市交通的治理之路——以山东省济南市为例	山东师范大学	沈可儿	202029000326	肖富茜/202029000306，邵焕芬/202029000108，王岳文乐/202129500122	韩庆龄



序号	项目名称	学院	项目类型	项目所属一级学科	是否重点支持领域项目	是否同意自筹	立项级别	主持人				
								学号	联系方式	性别	姓名（自动生成）	专业班级（自动生成）
305	蒺藜苜蓿对苯磺隆除草剂耐性基因的筛选研究	草业学院	创新训练项目	农学	是	是	学校资助项目	20200204363	13011498585	女	张紫鑫	草业(创新)2001
306	黄河三角洲盐碱地土壤碳排放观测系统的开发与利用	草业学院	创新训练项目	农学	是	是	学校资助项目	20200205137	13844555225	男	杨雨泽	草业(创新)2001
307	紫花苜蓿对邻生杂草荫蔽胁迫的生理生化响应	草业学院	创新训练项目	农学	是	是	学校资助项目	20200200086	19953600481	女	徐小燕	草业(创新)2001
308	狗尾草饲用价值评价及营养品质近红外体系建立	草业学院	创新训练项目	农学	是	是	学校资助项目	20200204890	18298340340	女	钟妮娜	草业(创新)2001
309	不同比例紫花苜蓿与无芒雀麦混播对茎叶功能性状的影响	草业学院	创新训练项目	农学	是	是	学校资助项目	20200200077	13812913121	女	刘馨月	草业(创新)2001
310	海滨雀稗对山东省滨海盐碱地土地改	草业学院	创新训练项目	农学	是	是	自筹经费项目	20200200081	15725211238	女	刘奕君	草业(创

七、学生工作

（一）青岛农业大学兼职辅导员（2022.10-2023.10）



聘书

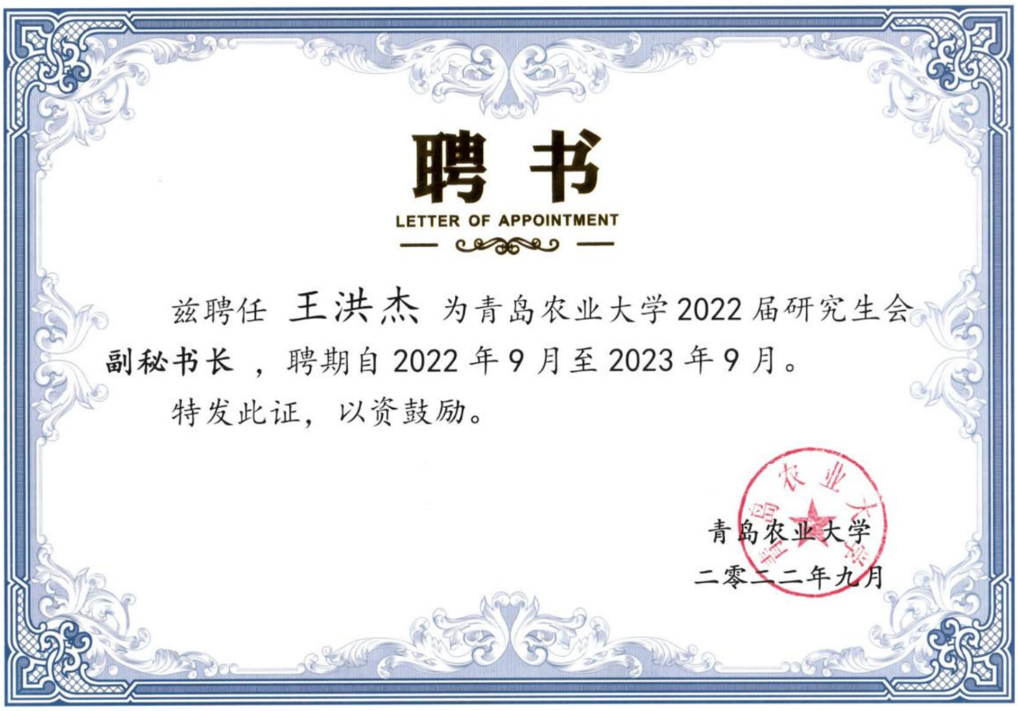
LETTER OF APPOINTMENT

兹聘任 王洪杰 为青岛农业大学兼职辅导员，聘期一年。

青岛农业大学

二〇二二年十月

（二）青岛农业大学研究生会副秘书长（2022.09-2023.09）



聘书

LETTER OF APPOINTMENT

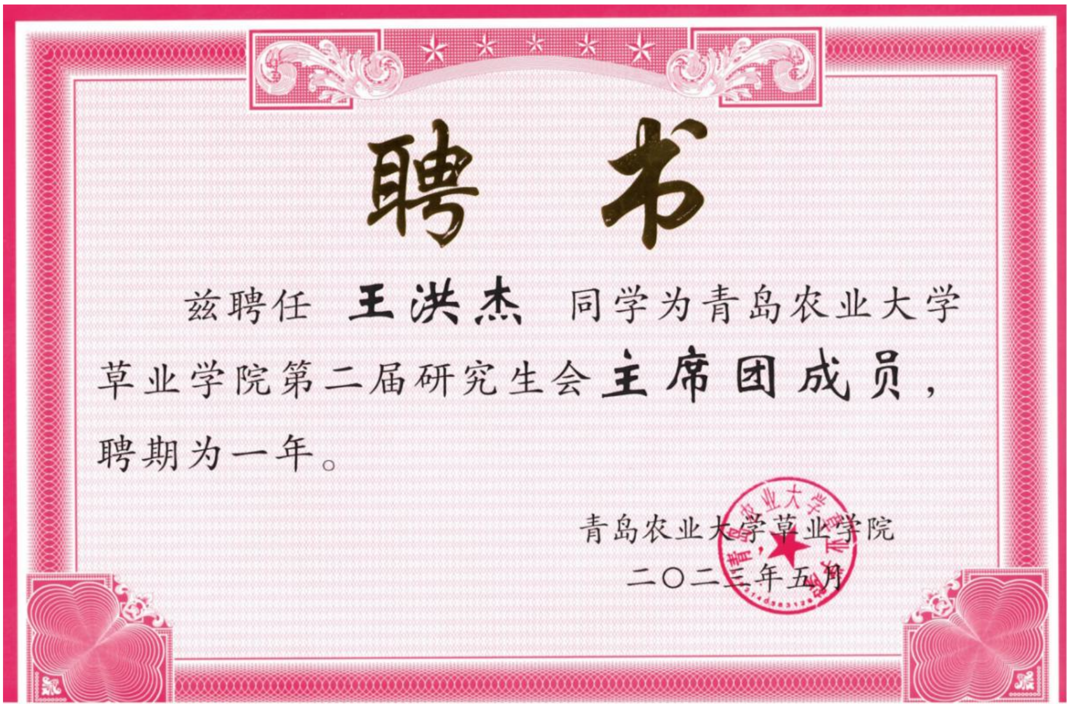
兹聘任 王洪杰 为青岛农业大学2022届研究生会副秘书长 ，聘期自2022年9月至2023年9月。

特发此证，以资鼓励。

青岛农业大学

二零二二年九月

（三）青岛农业大学草业学院主席团成员（2023.05-2024.05）

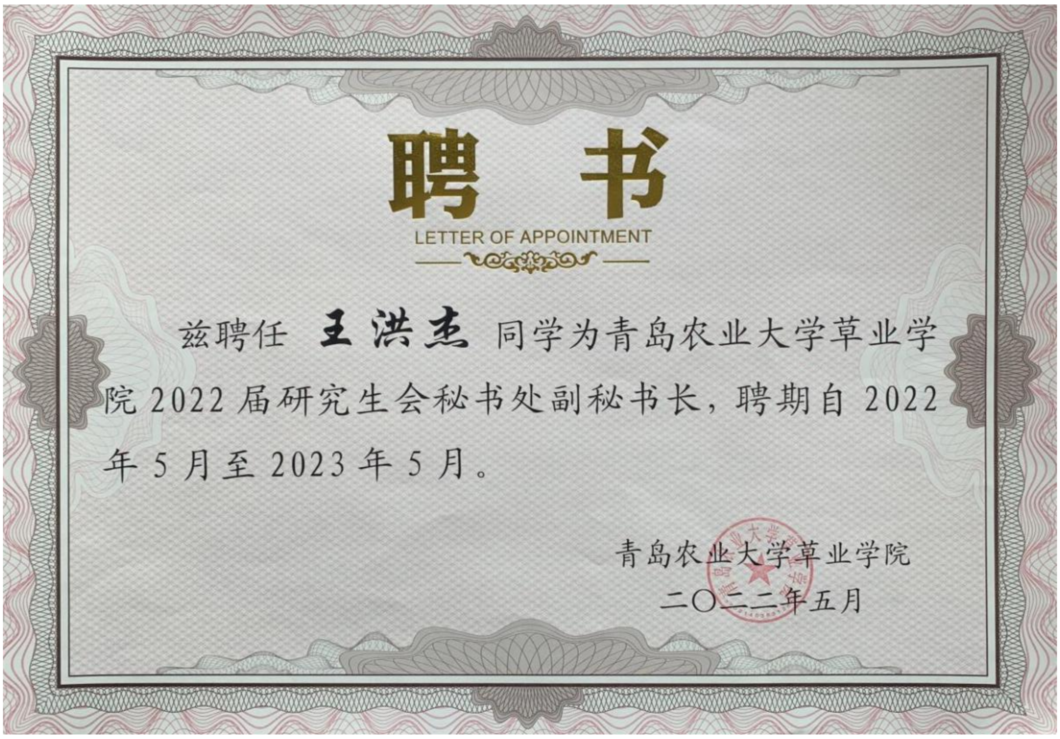


聘 书

兹聘任 **王洪杰** 同学为青岛农业大学草业学院第二届研究生会**主席团成员**，聘期为一年。

青岛农业大学草业学院
二〇二三年五月

（四）青岛农业大学草业学院副秘书长（2022.05-2023.05）



聘 书
LETTER OF APPOINTMENT

兹聘任 **王洪杰** 同学为青岛农业大学草业学院2022届研究生会秘书处副秘书长，聘期自2022年5月至2023年5月。

青岛农业大学草业学院
二〇二二年五月