Research Article

Journal of Emergency Medicine: Open Access

Whole-genome Sequencing and Analysis of *Apocynum Cannabinum*

Guoqi Li[1,2], Lixiao Song[1,2], Jinfeng Che[1,2], Yanyun Chen[1,2] and Juan Li[3]

[1]Breeding Base for State Key Laboratory of Land Degradation and Ecological Restoration in Northwest China, Ningxia University, Yinchuan, China

[2]Key Laboratory for Restoration and Recovery of Degraded Ecosystem in Northwest China of Ministry of Education, Ningxia University, Yinchuan, China

[3]North China University of Science and Technology, Tangshan,China

[*]Corresponding Author
Guoqi Li,
[1]Breeding Base for State Key Laboratory of Land Degradation and Ecological Restoration in Northwest China, Ningxia University, Yinchuan, China
[2]Key Laboratory for Restoration and Recovery of Degraded Ecosystem in Northwest China of Ministry of Education, Ningxia University, Yinchuan, China

Submitted: 2023, Aug 10; **Accepted**:2023, Aug 28: **Published**: 2023, Sep 12

Citation: Li, G., Song, L., Che, J., Chen, Y. ., Li, J. (2023). Whole-genome Sequencing and Analysis of Apocynum Cannabinum. *J Emerg Med OA, 1*(1), 43-52.

Abstract

Background

Apocynum cannabinum is an important plant resource from the Apocynaceae family. However, the lack of complete genome information has severely impeded research progress of molecular biology research in this plant. Whole-genome sequencing can provide an in-depth understanding of species growth, development, and evolutionary origin, and is the most effective method for scientifically exploring the ecological and economic value of a plant.

Methods and results

In this study, we employed Illumina HiSeq, single-molecule real-time sequencing, 10X genomics linked reads, and chromatin interaction (Hi-C), a new assembly technique, to successfully assemble the whole draft genome for A. cannabinum (260 Mb). The super-scaffold N50 genome size from the Hi-C assisted assembly was 21.16 Mb and was anchored to 11 chromosomes, resulting in a high-quality reference genome at the chromosome level (2n = 2x = 22). We further annotated, analyzed, and predicted 22,793 protein-coding genes, of which the functions of 95.6% were already annotated, 92.3% contained conserved protein domains, and 78.7% were aligned to known metabolic pathways.

Conclusion's:

This high-quality A. cannabinum genome can be used to analyze growth and development and evaluate gene evolution at the genome level, as well as assist in the comparative genomics and genetic modification of other important medicinal plants in Apocynaceae. Comparative analysis of the gene families showed that A. cannabinum speciated around 35.8 (27.0–46.9) million years ago.

Keywords Apocynum Cannabinum, Whole-Genome Sequencing, 10x Genomics Linked Reads, I-C, Speciation

1. Introduction

Apocynum venetum L is a perennial herb plant or small shrub that is widely distributed in saline-alkali soils of arid semi-arid area. It has a broad ecological distribution in Northwestern deserts in China and is a multi-purpose plant with ecological, economic, and medicinal value (1, 2). Because of its excellent ecological and economic value, *A. venetum* has a promising future of industrialization (3). To increase *Apocynum's* genetic diversities and ecological and economic value, our laboratory, with the support of project from the State Forestry Administration of the People's Republic of China, introduced *Apocynum cannabinum* from the abroad in 2004 and successfully cultivated the species in 2006. Preliminary studies showed that *A. cannabinum* has better drought resistance, thicker and straighter stalks, larger leaves, and fewer bifurcations than A. venetum (4). *A. cannabinum* is also known as dogbane or prairie dogbane and it is mostly produced in temperate and subtropical regions in North America (5). *A. cannabinum* has strong adaptability and has high early succession rate (6,7). Besides traditional fibers, *A. cannabinu*m also contains cardiac glycosides, resin, volatile oils, rubber, tannin, and starch

(8). Native American traditional medicine also uses A. cannabinum roots to treat heart disease and ascites (9). Apocynum is a genus of Apocynaceae. There are 44-46 genera and 145-176 species of Apocyaceae in China. They are mainly distributed in provinces South of the Yangtze River and coastal islands, and a few in the North and Northwest (10). Because Apocyaceae plants represent unique evolutionary groups and their important economic value, the systematic status of oleander family and the relationship between families and genera have been the research focus of phylogenetic scientist (11).

With the development and maturity of sequencing technology, an increasing number of plant genomics studies had been initiated to elaborate the molecular mechanism of plant growth, development, evolution and origin at the genome level (12). The second-generation and third-generation sequencing, containing 10X genomics linked reads and Hi-C, made many complex plant genomes sequencing possible (13). The technology of 10X genomics linked reads is essentially to introduce the barcode sequence into the long sequence segment, and then used them to construct more accurate super-scaffolds to explore more detailed genetic information (14). Hi-C sequencing method divides second-, third- generation or optical map-assisted assembled draft genome sequences into chromosome sets to determine the order and orientation of the various sequences on the chromosomes (15). Therefore, the combination of 10X genomics linked reads and Hi-C three-dimensional conformation capture technology greatly increases the accuracy of whole-genome sequencing and assembly quality. Although there are a lot of whole sequencing of economic plants and crop plants have been finished, the sequencing data about A. cannabinum are insufficient relatively, which hinder the further study of Apocynum, especially in pharmacology, molecular biology and genetic evolution. To evaluate the genome composition, component analysis, metabolic pathways, phylogenetic relationships, and speciation time of A. cannabinum, we carried out whole-genome sequencing of A. cannabinum, which will be helpful to elucidate its stress resistance mechanisms at the molecular level ultimately. Our findings provide theoretical support for the subsequent screening of superior genes, variety improvement, genetic engineering, and the scientific utilization of A. cannabinum.

2. Materials and Methods

2.1Experimental Materials

The experimental material was cultivated in pots after the belowground roots had been collected from the A. cannabinum experimental base in Pingluo county, Shizuishan city, Ningxia Province in end-March 2021. In mid-April 2021, healthy plants showing good growth were selected, and young leaves and stems at the apex were collected, snap-frozen in liquid nitrogen, and stored in a –80°C freezer for subsequent experiments.

2.2 Experimental Methods

2.2.1 Sample Extraction and Measurements

Whole-genome sequencing requires high DNA integrity and purity, and thus the modified cetyltrimethylammonium bromide (CTAB) method was used to extract genomic DNA from A. cannabinum (16). Agarose gel electrophoresis and NanoDrop 2000 were used to measure the integrity and concentration of the template, respectively, with the requirement of A260/280 ≥1.80.

2.2.2 Library Construction and Sequencing

Qualified DNA samples were fragmented using a Covaris ultrasonicator. Following this, magnetic beads were used for the enrichment and purification of large DNA fragments. This was followed by end terminal repair, addition of A-tails, addition of sequencing adapters, purification, and PCR amplification. We constructed two libraries of PacBio, 10X genomics linked reads, Hi-C and second-generation small fragment libraries, respectively , and sequenced by Illumina Hiseq and PacBio RSII sequencing platforms.

2.3 Genome Assembly

2.3.1 PacBio Genome Assembly

First, self-correction of the PacBio data was performed whereby the reads were aligned with each other. Self-correction was conducted according to the insertion and deletion of bases and the probability of sequencing errors to obtain pre-assembly reads. Following this, data assembly was carried out. As third-generation data reads are long (mean read length: 10–15 kb, longest read >40 kb), the Overlap-Layout-Consensus (17) method was used, i.e., the overlapping relationship of the reads was used for splicing to obtain a consensus sequence. Finally, the software Pilon (18) was used for another round of correction using the second-generation data of the assembly results from the previous step in order to increase the accuracy of the results, ultimately obtaining high-quality consensus sequences.

2.3.2 10x Genomics-Assisted Genome Assembly

In order to obtain high-quality genome assembly sequences, 10X Genomics-assisted assembly was also used in addition to the third-generation PacBio data assembly, mainly using fragscaff (https://sourceforge.net/projects/fragscaff/) software. Linked reads obtained from the 10X genomics linked reads library sequencing were used for alignment with the consensus sequence obtained from the third-generation assembly results. Linked reads were assembled on the original basis to form scaffolding (19).

2.3.3 Hi-C-Assisted Genome Assembly

First, quality control and adapter removal were performed on the generated data to obtain high-quality clean reads. Following this, data were aligned to the assembled genome sequence, and PCR repeats were removed. The interaction data obtained after noise correction were used to construct a chromosome interaction matrix. Based on the structural characteristics of the chromosomes in three-dimensional space, suitable clustering models were selected to anchor non-localized scaffolds on the chromosomes, and the correct order and direction of the scaffolds were determined by the corresponding sorting algorithm, and the whole genome sequence at the chromosome level was assembled with the software

LACHESIS(https://sourceforge.net/projects/lachesisproject/

2.3.4 Evaluation of A. Cannabinum Genome Assembly Quality

CEGMA (Core Eukaryotic Genes Mapping Approach) and BUSCO (Benchmarking Universal Single-Copy Orthologs) (20) were used to evaluate the assembly quality of the A. cannabinum genome. CEGMA evaluation selects conserved genes (248 genes) from six eukaryotic model organisms to construct a core gene library and uses tblastn, genewise, and geneid for genome evaluation. In BUSCO, 1440 orthologous single-copy genes from the plant database were used to evaluate the integrity of the assembled genome. To evaluate the assembly accuracy, small fragment library reads were aligned to the assembled genome using BWA software (21), and the distribution of the alignment rate, genome coverage, and depth of coverage of the reads were calculated to assess assembly integrity and sequencing uniformity.

2.4 Genome Annotation Analysis

2.4.1 Repeat Sequence Annotation

The default parameters of LTR_FINDER (22) (http://tlife.fudan.edu.cn/ltr_finder/; with -C -w 2 -s), RepeatScout (23) (http://www.repeatmasker.org/;with-sequence-freq-output), and RepeatModeler (24) (http://www.repeatmasker.org/RepeatModeler.html; with-database-engine ncbi-pa 15) were used to construct a de novo repeat sequence database of the A. cannabinum genome. The homologous repeat sequence database RepBase (25) (http://www.girinst.org/repbase/) was then used for integration, and RepeatMasker (26) (http://www.repeatmasker.org/) was used for repeat sequence annotation of the A. cannabinum genome.

2.4.2 Annotation of Non-Coding RNA

The tRNAscan-SE (27) (http://lowelab.ucsc.edu/tRNAscan-SE/) software was used to identify tRNA sequences in the genome based on the structural characteristics of the tRNAs. As rRNAs are highly conserved, the rRNA sequences of phylogenetically related species were used as reference sequences, and blast alignment was used to identify rRNAs in the genome. Rfam covariance models were used along with the INFERNAL (28) (http://infernal.janelia.org/) in Rfam to predict miRNA and snRNA sequence information in the A. cannabinum genome.

2.4.3 Gene Structure Prediction

First, the repeat sequence mask obtained from the annotation by RepeatMasker (with -nolow -no_is -norna -parallel 1 -lib -s) was used for de novo prediction of the A. cannabinum genome using Augustus (29) (http://bioinf.uni-greifswald.de/augustus/; with--species=pasa1-uniqueGeneId=true--noInFrameStop=true--gff3=on--genemodel=complete--strand=both) and GlimmerHMM (30) (http://ccb.jhu.edu/software/glimmerhmm/). Blast (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to align the protein sequences of four phylogenetically related species, namely crown flower (Calotropis gigantea), Madagascar periwinkle (Catharanthus roseus), Arabian coffee (Coffea arabica), and flax (Linum usitatissimum), to the A. cannabinum genome to predict the structure sets of homologous genes. Following this, blat (31) (http://genome.ucsc.edu/cgi-bin/hgBlat) was used for the alignment of the expressed sequence tag (EST) data to predict gene structures. Finally, EVidenceModeler (24) (http://evidencemodeler.sourceforge.net/) was used in combination with the transcriptome alignment data to combine the gene sets predicted by the different methods into a non-redundant and intact gene set.

2.4.4 Functional Annotation of the Genes

The gene sets obtained from the prediction of A. cannabinum gene structures were aligned using the Swiss-Prot (32) (http://www.uniprot.org/) and Non-redundant (Nr) (33) (https://academic.oup.com/nar/article/33/suppl_1/D121/2505359?login=false) databases using alignment software, and the best alignment result was used to determine gene function. The A. cannabinum genome was then aligned with Kyoto Encyclopedia of Genes and Genomes (KEGG) (34) metabolic pathways to determine the functions of the genes in the metabolic network and signaling pathways. In addition, InterPro (35) (https:// www.ebi.ac.uk/interpro/) software was used for Gene Ontology (GO) (36) annotation and prediction of gene motifs and domains.

2.5 Comparative genome analysis

In this study, we selected the genomes of 15 plants for homologous gene analysis. These genomes included *C. roseus, C. gigantea, Medigo truncatula, Populus trichocarpa, L. usitatissimum, Gossypium raimondii, Gossypium hirsutumD, Gossypium hirsutumA, Gossypium barbadenseA, Morus notabilis, Arabidopsis thaliana, Glycyrrhiza uralensis, Gossypium barbadenseD, Phyllostachys heterocycla, and Corchorus capsularis.* The CDS sequences of the aforementioned plants were obtained from various databases (NCBI, Ensembl, Phytozome). Before analysis, redundancy was removed from all data and only the longest protein sequence was retained for each gene. OrthoMCL (37) was used for homologous gene analysis. MUSCLE (38) (http://www.drive5.com/muscle/) alignment was carried out on various families by using the sequences of all single-copy genes. RaxML (39) (http://sco.h-its.org/exelixis/web/software/ raxml/index.html) was used for the construction of phylogenetic trees based on the sequence alignment results using the maximum likelihood method (ML TREE). Finally, CAFÉ (40) (http://sourceforge.net/projects/cafehahnlab/) was used for gene family expansion and contraction analysis.

3. Results

3.1 Sequencing Data Statistics

The PacBio platform were used for whole genome sequencing the A. cannabinum genome. Table1 represents total sequencing volume was 117.13 G, and the coverage was 490.04 X based on calculations of the genome size (239.02 M) estimated from a survey (41). In addition, the Illumina platform was used for the sequencing of a constructed second-generation small fragment library.

Pair-end libraries	Insert size	Total data (G)	Read length (bp)	Sequence coverage (X)
llumina reads	350 bp	31.94	150	133.63
PacBio reads	29.8 kb	8.76	11734	162.16
10X_Genomics	500–700 bp	46.43	150	194.25
Total	-	117.13	-	490.04

Table 1: Apocynum cannabinum genome sequencing data statistics

3.2 Assembly Results and Statistics

The genome size was 260 Mb, the contig N50 was 3.11 Mb, and the scaffold N50 was 4.19 Mb (scaffolds of 100 bp and above were selected for assembly results). The GC analysis showed that the GC content of *A. cannabinum* was 33.26%. Through Hi-C-assisted assembly, the final genome sequence was anchored to 11 chromosomes. The draft genome scaffold chromosome anchoring results showed that the constructed super-scaffold N50 was 21.16 Mb and the number of super-scaffolds was six, which was significantly decreased. The number of scattered scaffolds was low, showing good assembly results. Among the scaffolds, N90 was 11, indicating that 90% of sequences were on the 11 chromosome. A comparison of N50, N60, N70, N80, and N90 showed that the Hi-C assembled *A. cannabinum* genome result was significantly improved from the original draft genome (Table 2).

Sample ID		Length		Number	
		Contig (bp)	Scaffold (bp)	Contig (bp)	Scaffold (bp)
10X_Genomics	Total	259,856,155	260,145,681	479	439
	Max	9,567,299	15,154,524	-	-
	Number>=2000	-	-	479	439
	N50	3,107,135	4,189,930	26	19
	N60	2,442,000	3,120,689	35	27
	N70	1,722,976	2,442,000	48	36
	N80	1,116,808	1,547,882	66	49
	N90	257,161	339,954	109	80
Hi-C	Total	259,856,155	260,165,581	569	330
	Max	9,567,299	26,508,341	-	-
	Number>=2000	-	-	567	330
	N50	2,442,000	21,162,547	33	6
	N60	1,938,819	20,272,879	45	7
	N70	1,256,579	19,617,059	61	9
	N80	786,196	19,407,958	87	10
	N90	207,482	17,325,238	152	11

Table 2: *A. cannabinum* genome assembly results

A heatmap was used to validate the Hi-C assembly results (Fig. 1). From the heatmap, we can see that the interaction relationships near the diagonal were far stronger than those far away from the diagonal. The overall results showed that there was no significant clustering error between the *A. cannabinum* chromosomes.

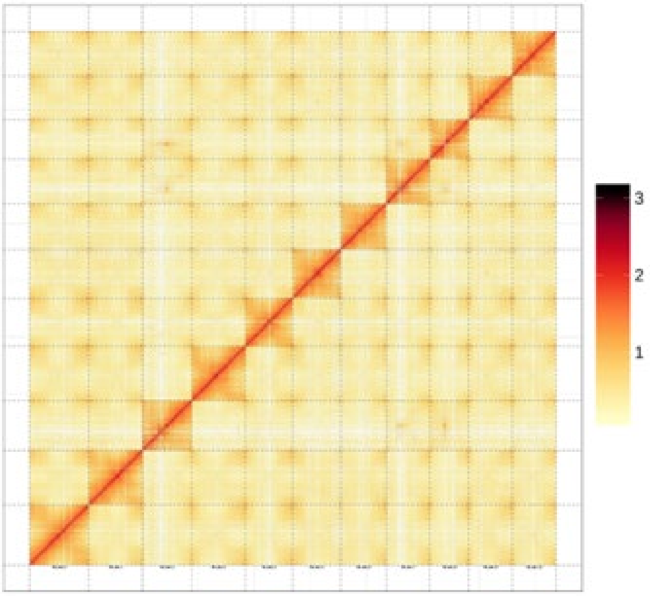

Figure 1: *A. cannabinum* chromosome interaction heatmap

3.3 Evaluation of Assembly Results

BUSCO and CEGMA software were used to evaluate the assembly results for A. cannabinum. The results showed that of the 1440 orthologous single-copy genes in the plant database, 1348 genes (C=93.6%) were completely covered, of which 1258 single copies were covered (S=87.4%), 89 multiple copies were covered (D=6.2%), 30 genes were not completely covered (F=2.1%), and 62 genes were missing (M=4.3%). This suggests that the vast majority of single-copy genes was completely assembled, that there was no over-assembly or re-assembly, and that the assembly accuracy and integrity were good. At the same time, 227 genes were assembled from 248 core eukaryotic genes (CEGs), indicating that the assembly results were relatively complete. Alignment using BWA software showed that the alignment rate of all small fragment reads to the genome was 94.80% and the coverage was 96.95%, demonstrating that the reads and the assembled genome possess good consistency.

4. Repeat Sequence Annotation

Repeat sequences can be classified as tandem repeats and interspersed repeats. Tandem repeats include microsatellite sequences and minisatellite sequences. Interspersed repeats are also known as transposon elements and include DNA transposons that use DNA-DNA transposition and retrotransposons. Commonly seen retrotransposons include long terminal repeats (LTRs), long interspersed elements (LINEs), and short interspersed elements (SINEs). Around 36.62% of sequences in the A. cannabinum genome were identified as repeat sequences, of which DNA transposons, LTRs, LINEs, and SINEs accounted for 2.22%, 28.45%, 1.01%, and 0%, respectively (Table 3).

	Denovo+Repbase length (bp)	% in genome	TE protein length (bp)	% in genome	Combined TE length (bp)	Type
DNA	5,282,438	2.03	765,646	0.29	5,763,303	2.22
LINE	1,744,782	0.67	1,157,517	0.44	2,630,319	1.01
SINE	1,597	0.00	0	0	1,597	0.00
LTR	72,896,704	28.02	16,110,716	6.19	74,004,538	28.45
Unknown	14,826,765	5.70	0	0	14,826,765	5.70
Total	92,189,164	35.44	18,033,358	6.93	93,170,210	35.81

Note: Denovo+Repbase are transposon elements obtained after the databases predicted by RepeatModeler, RepeatScout, Piler, and LTR_FINDER were combined with the RepBase nucleic acid database. This was followed by integration using Uclust software according to the 80-80-80 principle, following which RepeatMasker was used for genome annotation. The TE proteins are transposon elements that were obtained when the RepeatProteinMask software was used to annotate genomes in the RepBase protein database. Combined TEs is the result obtained by combining the results obtained from these two methods and following the removal of redundant genes. Unknown indicates that this repeat sequence could not be classified by RepeatMasker.

Table 3: Statistics of repeat sequence classification results for *A. cannabinum*

4.1 Gene Structure Annotation

Through the annotation of the A. cannabinum genome, a total of 22,793 non-redundant protein coding genes were predicted. The mean gene sequence length of the gene was 3532 bp, the mean length of coding sequence was 1206 bp, and the exons and introns were 231 bp and 551 bp, respectively. The mean number of exons in every gene was 5.22. Comparison of the lengths of genes, coding sequences, introns, exons, and the number of exons in A. cannabinum with C. gigantea, Catharanthus roseus, C. arabica, and L. usitatissimum showed that the lengths of the coding sequences and exons and the number of exons in every gene in A. cannabinum were closest to C. gigantea and L. usitatissimum, while C. roseus had lower numbers than these three species and C. arabica had higher numbers than these species. The lengths of the genes and introns of A. cannabinum were most similar to C. arabica, while the lengths of the genes and introns of L. usitatissimum were significantly lower than the other four plant species. The lengths of almost all introns in the five plant species were within 1000 bp, and introns with lengths greater than 2000 bp were extremely rare or absent (Fig. 2).

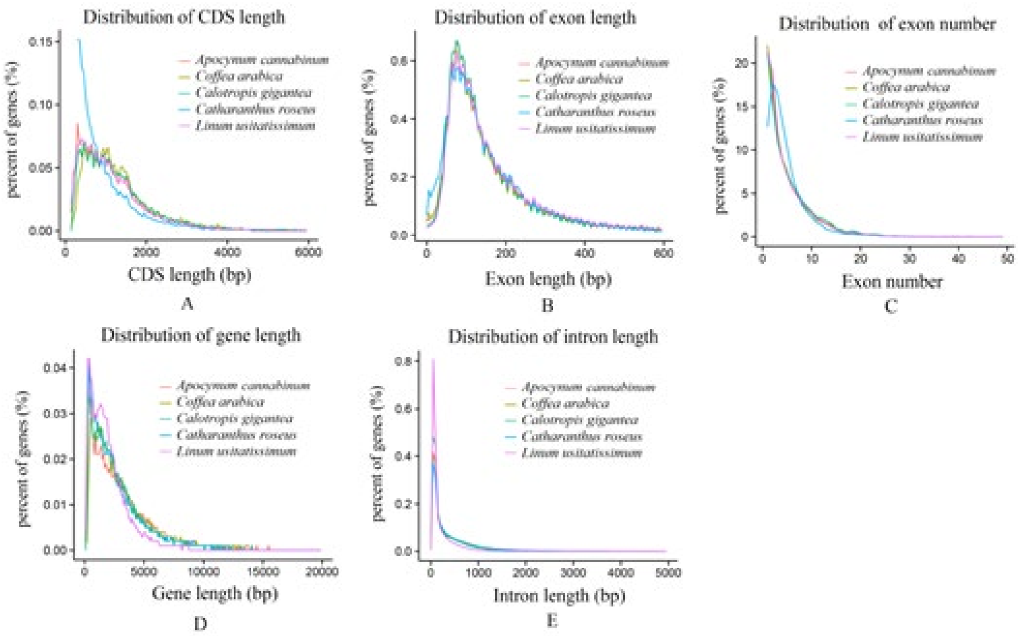

Figure 2: Comparison of various elements in phylogenetically related species

4.2 Annotation of Gene Function and Non-Coding Rnas

A total of 78.70% of genes could be aligned to known metabolic pathways; 12,823 functional genes could be annotated by GO function; and 4.4% of genes had unknown functions. Among the annotated protein-coding genes, 21,780 genes had homologous sequences in the protein database, accounting for 95.6% of annotated genes. Furthermore, 95.2% of genes were aligned to non-redundant protein databases, of which 92.30% contained conserved protein domains, showing that the functions of most A. cannabinum genes were relatively conserved. In addition, 689 tRNAs, 3934 rRNAs, 592 snRNAs, and 159 miRNAs were annotated in the A. cannabinum genome (Fig.3).

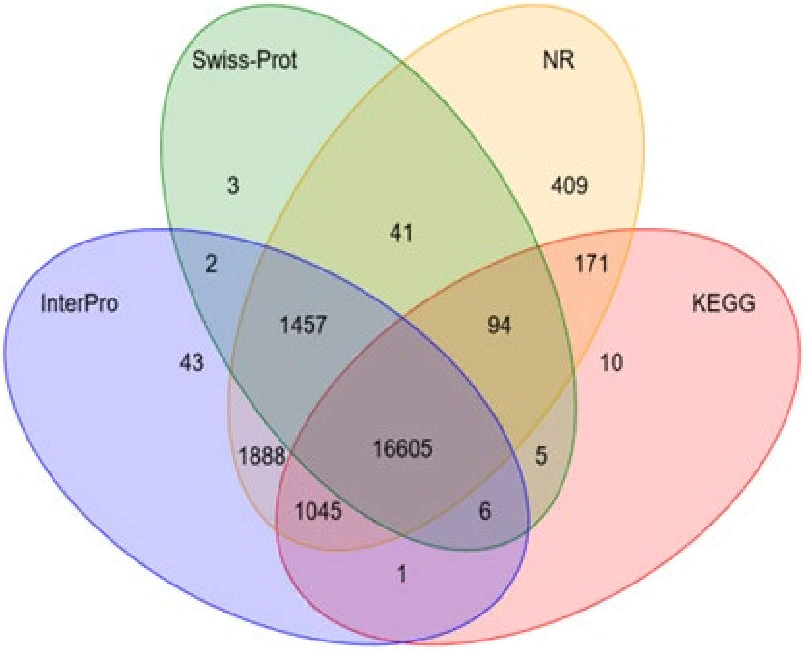

Figure 3: Gene function annotation results for A. cannabinum

4.3 Comparative Genomic Analysis of A. cannabinum

Clustering analysis of the gene families indicated that 43,837 gene families were clustered in 17 species, of which 3084 gene families were common to all these species, while there were 55 common single-copy gene families in the different species. The clustering results of A. cannabinum, A. venetum, C. roseus, C. gigantea, and C. capsularis were used to plot a Venn Diagram (Fig. 4).

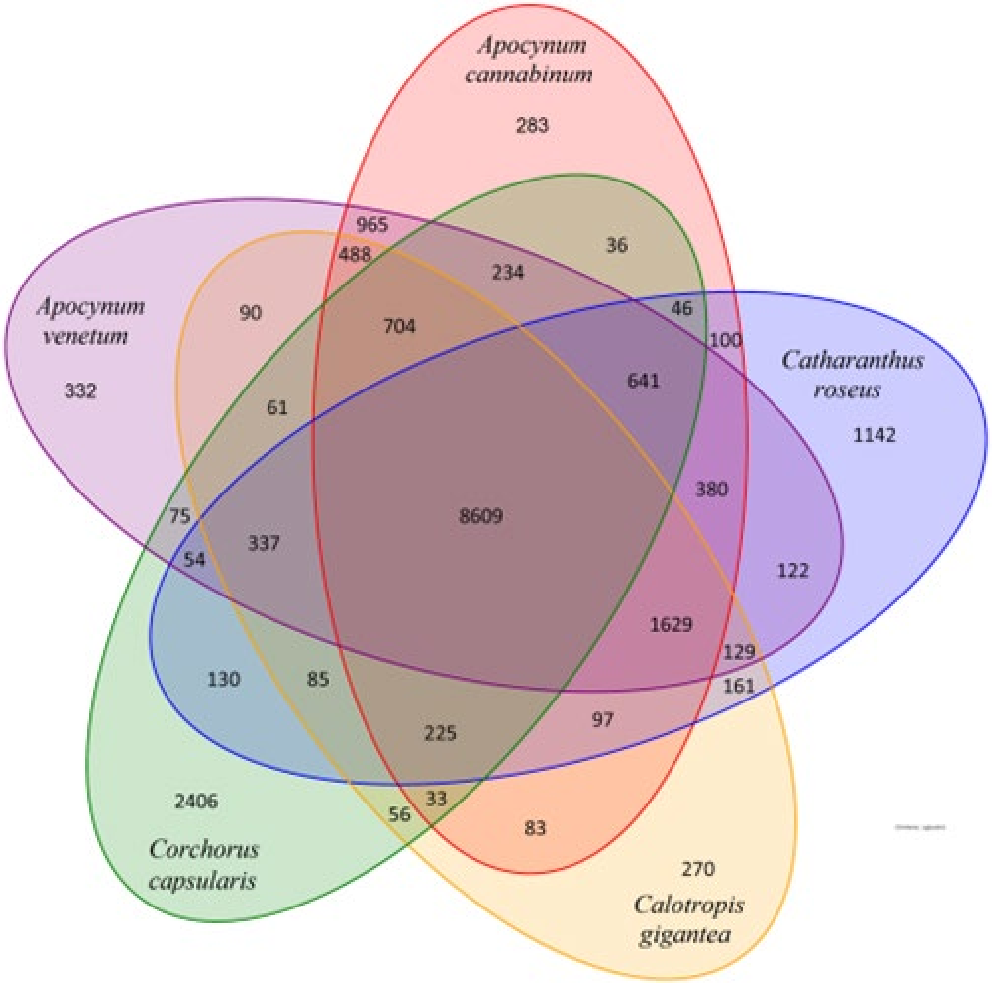

Figure 4: Annotation: overlapping areas in the circles show gene families that are common between species; the numbers represent the number of gene families. Non-overlapping areas show the number of gene families that are unique to that species. The sum of numbers within the entire circle represents the total number of gene families in that species.

From the Venn diagram, we can see that A. cannabinum contains 283 species-specific gene families and 449 genes compared with other species. The gene family expansion and contraction analysis results showed that the most recent common ancestor (MRCA) contained 43,823 gene families, of which 399 gene families and 166 genes were expanded in A. cannabinum, while 492 gene families and 231 genes were contracted (Fig. 5).

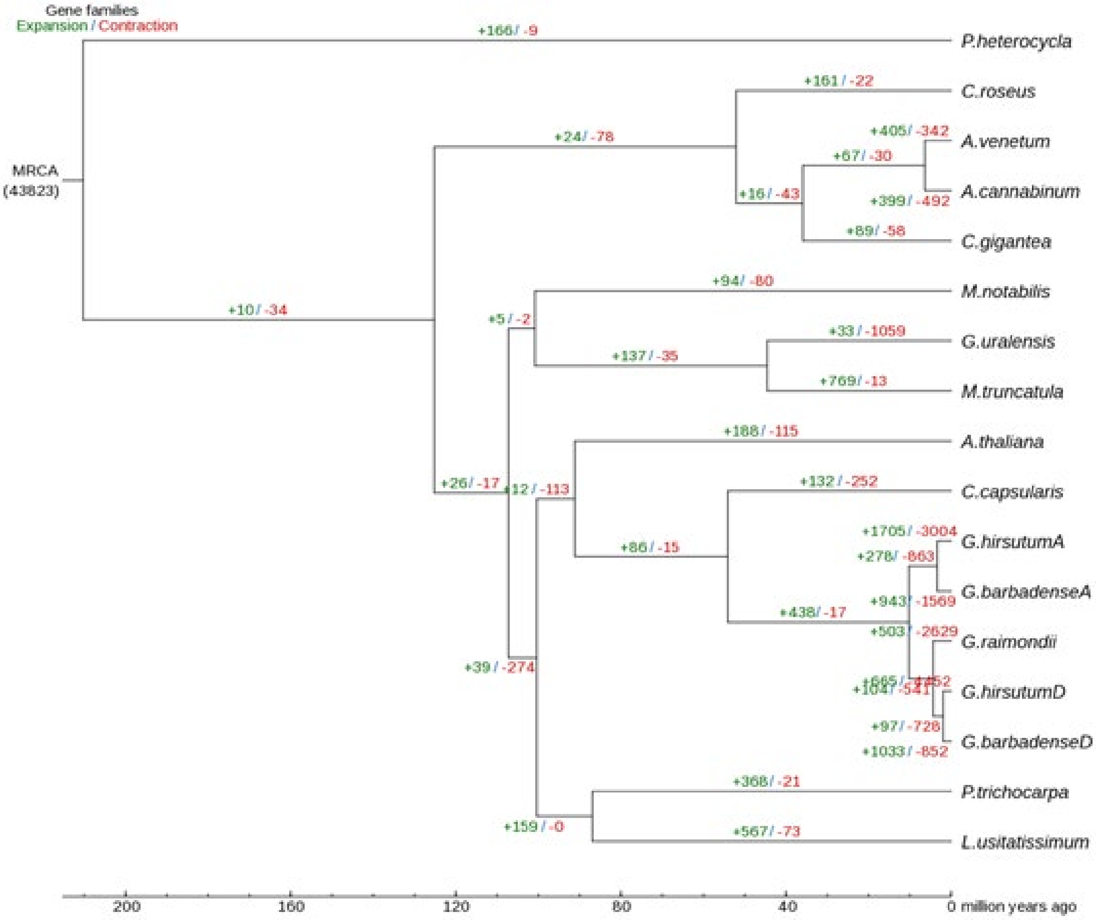

Figure 5: Annotation: as shown in the above figure, green numbers represent the number of gene families that expanded during evolution, and the red numbers represent the number of gene families that contracted during evolution.

The gene family clustering analysis results showed that C. roseus, A. venetum, A. cannabinum, and C. gigantea clustered in the same taxon. The Markov chain Monte Carlo (MCMC) program in PAML software indicated that the speciation time for A. cannabinum was around 35.9 (26.7–46.8) million years ago.

5. Discussion

In this study, we employed Illumina HiSeq, single-molecule real-time sequencing, and 10X genomics as well as chromatin interaction (Hi-C), a new assembly technique, to successfully assemble the whole draft genome for A. cannabinum for the first time. The whole-genome sequence of A. cannabinum was 260 Mb, the size of the genome super-scaffold N50 was 21.16 Mb, and the scaffold N50 was 4.19 Mb. The combination of the Hi-C sequencing results and high-density genetic maps resulted in the scaffolds being anchored to 11 chromosomes. The super-scaffold N50 length was 21.16 Mb, which is currently the best result among all assembled genomes from the Apocynaceae family. There are four species genomes of Apocynaceae family (Catharanthus roseus, Calotropis gigantean, Rhazya stricta and Asclepias syriaca) have been sequenced and published. However, since the contig N50 of C. roseus and A. syriaca is all less than 10 kb, it is easy to cause fairly big errors in genome annotation and gene localization. In addition, the published genomes from the Apocynaceae family are at the chromosome level, while the present assembled A. cannabinum genome exceeds the chromosome level. Gene annotation refers to the specific role of biomolecules in the process of life. Genome annotation is an essential link connecting the sequence information of genes with specific biological processes. Therefore, the functional annotation process of all genes in the whole genome sequence of plants is also considered as the "metabolic reconstruction" of plants. The purpose of metabolic reconstruction is to identify metabolic pathways and which genes have this function in plants. However, the plant genome sequence contains a large number of repetitive regions, pseudogenes, as well as many new protein-coding and non-coding genes. The repetitive sequence can even be as high as 80% in plant genome. Although the gene duplication plays an important role in maintaining the regulation of gene expression, the spatial structure of chromosomes and genetic recombination, etc., it brings out many false positives in BLAST results, which increases the pressure of gene structure prediction and may lead to the high error rate of gene annotation. This provides a foundation for future studies on A. cannabinum. Gene annotation in A. cannabinum revealed a total of 22,793 non-redundant protein-coding genes, of which 21,780 have homologous sequences in the protein database, accounting for 95.6% of annotated genes. This prediction result is similar to the number of genes (21,164) in Rhazya stricta from the Apocynaceae family (42). Conversely, the number of genes in Asclepias syriaca (14,474) (43) and C. gigantea (18,197) (44) is lower than in A. cannabinum. A previous study also showed that no direct relationship exists between genome size and the number of genes in plants (45). The reasons for the large differences in the number of genes include: the selection of overly high threshold values for annotation, resulting in marginal data being excluded; the selection of only a few major public databases, resulting in incomplete coverage; a lack of homology annotation information for certain specific genes, which require functional validation in the future; and excessive structural annotation, resulting in the annotation of false positive genes (45). According to our study, A. annabinum has better drought tolerance than A. venetum (1, 4), we believe that it will be well explained when the genome is analyzed in the future.

Comparative genomic analysis of A. cannabinum indicated that a total of 43,837 gene families were clustered, of which 3084 were common gene families, 399 gene families and 1666 genes were expanded, and 492 gene families and 231 genes were contracted. This may be due to the varying degrees of acquisition and loss of genes (families) of each species during evolution (46). Additionally, A. cannabinum speciated around 35.9 (26.7–46.8) million years ago. Gene family clustering found that C. roseus, A. venetum, A. cannabinum, and C. gigantea were clustered in the same taxon, which is highly consistent with the angiosperm classification system IV (APG IV).

Conclusions

Whole-genome sequencing of A. cannabinum was carried out, and 117.13 G of raw data was obtained. The sequencing depth was 490.04 X, the assembled genome size was 260 Mb, and the contig N50 was 3.11 Mb. Multiple methods were used for the evaluation of the assembled versions. The results showed that A. cannabinum/s genome version had high consistency, integrity, and accuracy. A total of 22,793 protein-coding genes were predicted by genome annotation, and the ratio of repeat sequences was 35.81%, of which functions were predicted for 21,780 genes (95.6%). Clustering analysis of the 17 species was used to obtain 43,837 gene families, of which 3084 constituted common gene families. Comparative analysis of the gene families showed that A. cannabinum speciated around 35.8 (27.0–46.9) million years ago.

Acknowledgments This work is supported by Ningxia Key Research and Develoment Project (grant nunber: 2022BEG03006), and Ningxia Natural Science Foundation (grant nunber: 2020AAC03076). Especially, we heartily thank professor Dongmei Ma, Mr. Boxun Xie and Sheng Xie, they all gave us some useful revision advices.

Authors Contributions Guoqi Li conceived and designed the experiment, Lixiao Song and Jinfeng Che conducted experiment and data analysis, Guoqi Li, Lixiao Song, Jinfeng Che, Yanyun Chen and Juan Li wrote, discussed and rewrote the manuscript together.

Conflict of interest The authors declared that they have no competing interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Data availability All research data are available upon request.

References

1. Li, G. Q., & Chen, Y. Y. (2012). Physioecology of Apocynum.
2. Xie, W., Zhang, X., Wang, T., & Hu, J. (2012). Botany, traditional uses, phytochemistry and pharmacology of Apocynum venetum L.(Luobuma): A review. Journal of Ethnopharmacology, 141(1), 1-8.
3. Li, G., Zhao, P., & Shao, W. (2019). Cash crop halophytes of China. Sabkha Ecosystems: Volume VI: Asia/Pacific, 497-504.
4. Wang, D., Li, G., & Wang, L. (2012). Daily dynamics of photosynthesis and water physiological characteristics of Apocynum venetum and A. cannabinum under drought stress. Acta Botanica Boreali-Occidentalia Sinica, 32(6), 1198-1205.
5. DiTommaso, A., Clements, D. R., Darbyshire, S. J., & Dauer, J. T. (2009). The biology of Canadian weeds. 143. Apocynum cannabinum L. Canadian Journal of Plant Science, 89(5), 977-992.
6. Keever, C. (1979). Mechanisms of plant succession on old fields of Lancaster County, Pennsylvania. Bulletin of the Torrey Botanical Club, 299-308.
7. Mulhouse, J. M., & Galatowitsch, S. M. (2003). Revegetation of prairie pothole wetlands in the mid-continental US: twelve years post-reflooding. Plant Ecology, 169, 143-159.
8. Leidy. Indian use of Apocynum cannabinum as a textile fiber. Proceedings of the Academy of Natural Sciences of Philadelphia1884, 36, 30-30.
9. Duprey, A. J. B. (1905). A CASE OF MITRAL INCOMPETENCY AND ASCITES TREATED WITH APOCYNUM CANNABINUM. The Lancet, 166(4283), 955-956.
10. Li, B.T, Chen, X. M. (1997) Comparative review on Apocynaceae in flora reipublicae popularis sinicae and flora of China. Guihaia, 17 (4): 299-305.
11. Li, P.T, Leeuwenberg AJM, Middleton DJ. Flora of China. Beijing: Science Press and St. Louis: Missouri Botanical Garden 1995, 143-88.
12. Goff SA, Ricke D, Lan TH, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science 2002, 296 (5565): 92-100.
13. Burton JN, Adey A, Patwardhan RP, et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol 2013, 31(12): 1119-25.
14. Eisenstein M. Startups use short-read data to expand long-read sequencing market. Nat Biotechnol 2015, 33(5): 433-35.
15. Mascher M, Gundlach H, Himmelbach A, et al. A chromosome conformation capture ordered sequence of the barley genome.

Nature 2017, 544(7651): 427-33.

16. Xu M, Sun Y, Li H. EST-SSRs development and paternity analysis for Liriodendron spp. New Forest 2010, 40(3): 361-82.
17. Liu HL. The whole genome sequencing and analyzing of Ginkgo biloba. Nanjing Forestry University, 2017 (in Chinese)
18. Hansen KD, Brenner SE, Dudoit S. Biases in illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res 2010, 38(12): 131-38.
19. Jarvis DE, Ho YS, Lightfoot DJ, et al. The genome of Chenopodium quinoa. Nature 2017, 542 (7641): 307-12.
20. Hu W, Hou Y, Zhang F, et al. A Chromatin conformation analysis technology—Hi-C and extracting of chromatin conformation information. Genomics and Applied Biology 2015, 34(11): 36-44.
21. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 2013,1303.
22. Zhao X, Hao W. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 2007, 35(Web Server issue), W265-8.
23. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics 2005, 21(suppl_1), 351-58.
24. Haas BJ, Salzberg SL, Zhu W, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to assemble spliced alignments. Genome Biol 2008, 9(1): R7.
25. Jurka J. Repbase update, a database and an electronic journal of repetitive elements. Trends Genet 2000, 16(9): 418-20.
26. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics 2009, Supplement 25: 4.10.1-4.10.14.
27. Lowe TM, Chan PP. tRNAscan-SE On-line: Search and Contextual Analysis of Transfer RNA Genes. Nucl Acids Res 2016, 44: W54-57.
28. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 2013, 29: 2933-35.
29. Stanke M, Steinkamp R, Waack S, et al. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 2004, 32(Web Server issue): 309-12.
30. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. Bioinformatics 2004, 20(16): 2878-79.
31. Kent WJ. BLAT-the BLAST-like alignment tool. Genome Res 2002, 12(4): 656-64.
32. Bairoch A. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000, 28(1): 45–48.
33. Griffiths-Jones S, Moxon S, Marshall M, et al. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 2005, 33: D121-24.
34. Ogata H, Goto S, Sato K, et al. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 1999, 27(1): 29-34.
35. Zdobnov EM, Apweiler R. InterProScan-an integration platform for the signature-recognition methods in InterPro. Bioinformatics 2001, 17(9): 847-848.
36. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000, 25(1): 25-29.
37. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 2003, 13(9): 2178-89.
38. Robert CE. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004, 32: 1792-1797.
39. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 2003, 52(5): 696-704.
40. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat. Rev. Genet 2012, 13(5): 303-14.
41. Song L, Li G, Jin C, et al. Whole genome sequencing and development of SSR markers in Apocynum cannabinum. Journal of Plant Genetic Resources 2019, 20(5): 1309-6. (in Chinese)
42. Sabir JSM, Jansen RK, Arasappan D, et al. The nuclear genome of Rhazya stricta and the evolution of alkaloid diversity in a medically relevant clade of Apocynaceae. Sci. Rep-UK 2016, 6(1): 33782.
43. Weitemier K, Straub S, Fishbein M, et al. A draft genome and transcriptome of common milkweed (Asclepias syriaca) as resources for evolutionary, ecological, and molecular studies in milkweeds and Apocynaceae. Peer J 2019, 7: e7649.
44. Hoopes GM, Hamilton JP, Kim J, et al. Genome assembly and annotation of the medicinal plant Calotropis gigantea, a producer of anticancer and Anti-malarial Cardenolides. G3-Genes Genom Genet 2017, 8(2): 385-91.
45. Kellner F, Kim J, Clavijo BJ, et al. Genome-guided investigation of plant natural product biosynthesis. The Plant J 2015, 82(4): 680-692.
46. Lang K, Bi S, Li F. Genome-wide analysis of expansion and contraction of gene families in parasitic wasps. Journal of Anhui Agricultural University 2018, 45(5): 945-50. (in Chinese).