Bioscience Reports (2020) **40** BSR20202709
https://doi.org/10.1042/BSR20202709

PORTLAND PRESS

Research Article

Genome survey sequencing of *Atractylodes lancea* and identification of its SSR markers

Tingyu Shan[1,*], Junxian Wu[1,*], Daqing Yu[1], Jin Xie[2], Qingying Fang[1], Liangping Zha[1,3] and Huasheng Peng[1,4]

[1]College of Pharmacy, Anhui University of Chinese Medicine, Hefei 230012, China; [2]College of Pharmacy, Anhui Medical University, Hefei 230032, China; [3]Institute of Conservation and Development of Traditional Chinese Medicine Resources, Anhui Academy of Chinese Medicine, Hefei 230012, China; [4]Research Unit of DAO-DI Herbs, Chinese Academy of Medical Sciences, Beijing 100700, China

Correspondence: Liangping Zha (zlp_ahtcm@126.com) or Huasheng Peng (hspeng@126.com)

Atractylodes lancea (Thunb.) DC. is a traditional Chinese medicine rich in sesquiterpenes that has been widely used in China and Japan for the treatment of viral infections. Despite its important pharmacological value, genomic information regarding *A. lancea* is currently unavailable. In the present study, the whole genome sequence of *A. lancea* was obtained using an Illumina sequencing platform. The results revealed an estimated genome size for *A. lancea* of 4,159.24 Mb, with 2.28% heterozygosity, and a repeat rate of 89.2%, all of which indicate a highly heterozygous genome. Based on the genomic data of *A. lancea,* 27,582 simple sequence repeat (SSR) markers were identified. The differences in representation among nucleotide repeat types were large, e.g., the mononucleotide repeat type was the most abundant (54.74%) while the pentanucleotide repeats were the least abundant (0.10%), and sequence motifs GA/TC (31.17%) and TTC/GAA (7.23%) were the most abundant among the dinucleotide and trinucleotide repeat motifs, respectively. A total of 93,434 genes matched known genes in common databases including 48,493 genes in the Gene Ontology (GO) database and 34,929 genes in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. This is the first report to sequence and characterize the whole genome of *A. lancea* and will provide a theoretical basis and reference for further genome-wide deep sequencing and SSR molecular marker development of *A. lancea*.

Introduction

Atractylodes lancea (Thunb.) DC., which belongs to the Asteraceae (Compositae) plant family, has been widely used as traditional medicine in many countries, including China and Japan. The main pharmacologically active components of *A. lancea* are present in the volatile oils and include atractylon, atractylodin, hinesol and β-eudesmol [1,2]. Previous pharmacological studies have demonstrated that the extract and chemical constituents of *A. lancea* have promising anti-inflammatory, anti-cancer, and anti-microbial effects and have been used as a remedy for gastritis and gastric ulcers [3–5]. *A. lancea* also has anti-influenza properties, as atractylon has been shown to effectively kill influenza A virus subtypes H3N2 and H5N1, and influenza B [6]. In addition, *A. lancea* also has immunomodulatory activity [7].

Despite its important medicinal value, the genetic information of *A. lancea* has remained largely unknown. The chromosome number and karyotype of *A. lancea* have been studied. The chromosome number of *A. lancea* is $2n = 24$, and the karyotype is 2B [8–10]. In addition, the cloning and expression analysis of *DXS* [11], *DXR* [12], *HMGR* [13] and *FPPS* [14] genes in *A. lancea* have been reported; however, the genome size and genome-wide sequencing of *A. lancea* have not been reported. The genomes of Compositae plants, such as *Carthamus tinctorius* [15] and *Artemisia annua* [16], have been sequenced, and these genome studies can provide a reference for the development of these medicinal plants.

Next-generation sequencing (NGS) technology is a relatively new methodology that can enable the identification of large numbers of simple sequence repeat (SSR) markers. This technology has dramatically

*These authors contributed equally to this work.

Received: 14 August 2020
Revised: 03 October 2020
Accepted: 06 October 2020

Accepted Manuscript online: 07 October 2020
Version of Record published: 28 October 2020

Bioscience Reports (2020) **40** BSR20202709
https://doi.org/10.1042/BSR20202709

increased the sequence output while also reducing time and cost [17,18]. Genomic survey analysis is a method of combining NGS technology with K-mer analysis to obtain species genome size, GC content, heterozygosity rate, and repetition rate. This method has been used to accurately predict the whole genome sizes of *Xanthoceras sorbifolium* [19], *Pennisetum purpureum* [20], *Pistacia vera* [21], and other plant species. In the present study, we used genomic survey analysis to study the genome of *A. lancea*. The aims of the present study were threefold; first, to estimate the genome size, GC content, and heterozygosity rates for *A. lancea*; second, to characterize the genome-wide SSRs in the *A. lancea* genome using a genome survey, and third, to perform Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and KOG pathway analyses to further elucidate the main biological functions of the genome survey.

Materials and methods

Plant material

Fresh *A. lancea* (Thunb.) DC. plants were collected in July 2018 from Nanjing city, in Jiangsu Province, China (Supplementary Figure S1). Fresh leaves at the apex of *A. lancea* plants were transported in liquid nitrogen to the laboratory where they were stored at −80 °C until needed in subsequent experiments. Samples were authenticated by Prof. Huasheng Peng (Anhui University of Chinese Medicine).

Genomic DNA extraction and detection

The genomic DNA was extracted from young leaves of *A. lancea* by means of an improved CTAB method [22]. Approximately 100 mg of fresh leaf samples were frozen in liquid nitrogen and ground into a fine powder with a mortar and pestle. A modified CTAB extraction buffer (2% CTAB, 100 mM Tris-HCl pH = 8.0, 1.4 M NaCl, 20 mM EDTA pH = 8.0, 2% β-mercaptoethanol and 3% polyvinylpyrrolidone) was added and the mixture incubated for 90 min at 65 °C. After incubation, an equal volume of chloroform: isoamyl alcohol (24:1) was added, mixed thoroughly, and centrifuged for 5 min at 12,000 rpm. After centrifugation, the supernatant was transferred to a fresh tube, and isopropanol was added to precipitate the DNA. The extracted DNA was stored in TE buffer (10 mM Tris-HCl, 1 mM EDTA) at −20 °C. DNA concentration was determined using a NanoDrop 2000 spectrophotometer (Thermo Scientific, Wilmington, Delaware, U.S.A.), and purity was verified using the ratios OD_{260}/OD_{280} and OD_{260}/OD_{230}. The integrity and quality of the extracted DNA was verified using 1% agarose gel electrophoresis. For quality control, the DNA samples were randomly selected and sent to independent services for DNA sequencing analysis.

Library construction and sequencing

The DNA extracted from *A. lancea* was frozen on dry ice and sent to the Beijing Genomics Institute Co., Ltd. (Shanghai, China) for library construction and sequencing. The DNA was sheared randomly into small fragments and sequencing was performed using an Illumina HiSeq 2500 sequencing platform (Illumina, CA, U.S.A.). The clean reads were obtained by clusters passing filter to remove low-quality bases and adapter contamination. The high-quality clean data obtained were used for subsequent analyses.

17-mer analysis

Before genome assembly, the genome size, heterozygosity rate, and repeated sequences of *A. lancea* were determined to be consistent with the estimated genome size based on K-mer analysis. K-mer spectrum reflects the heterozygosity of the genome. K-mer analysis of *A. lancea* genome was conducted using *jellyfish* (version 2.1.4). The software *GenomeScope* (http://qb.cshl.edu/genomescope/) was used to fit the 17-mer spectrum and estimate the genome size.

Heterozygosity prediction and GC content analysis

De novo assembly of the genome was performed using *SOAPdenovo* software (version 2.04). Contigs and scaffolds were constructed to obtain the original genome sequence. The average depth and GC content of each window were calculated to produce the GC-depth plot and repeat content of the genome was determined according to the stratification of GC clusters.

Table 1 Statistical data from the 17-mer analysis

	K-mer number	Genome size (Mp)	Repeat (%)	Heterozygous ratio (%)	Used bases (Gp)	Used sequence depth (X)
17	223162157589	4159	89.2	2.28	251.12	60

Preliminary assembly of genome

Preliminary genome assembly and stitching was performed using *SOAPdenovo* software. K-mers were counted for each sequence read, and the frequency of each K-mer was determined. The assembled genome sequences were compared with the original data to determine GC content, contig coverage depth, length, and quantitative distribution of the assembled sequences.

SSR identification

MISA software (version 1.0, http://pgrc.ipk-gatersleben.de/misa/) was used to identify SSR markers. The criteria for identifying SSR motifs were as follows: the minimum number of nucleotide repeats was ten for mononucleotide repeats, six for dinucleotide repeats, and five for trinucleotide, tetranucleotide, pentanucleotide, or hexanucleotide repeat motifs.

Gene prediction and annotation

Clean reads were processed using *Trinity* software (https://github.com/trinityrnaseq/trinityrnaseq/wiki) to obtain the high-quality unigene library. The obtained unigene was analyzed using bioinformatics, including functional annotation and classification. The BLASTx comparison tool was used to compare the unigene with the protein database (E-value $\leq 1e^{-5}$). Functional annotation was based on similarity of the gene to the functional annotation information for the unigene encoded protein. The protein databases included NR (non-redundant protein database), COG (Cluster of Orthologous Groups), GO (Gene Ontology database), and KEGG (Tokyo Encyclopedia of Genes and Genomes).

Results

Sequencing data statistics

The Illumina HiSeq™ 2500 platform was used for high-throughput, paired-end sequencing to obtain 256.35 Gb of *A. lancea* raw bases. *SOAPnuke* software (version 1.6.5) was used to filter the original sequencing data, and a total of 251.12 Gb clean reads were generated after filtering out low-quality reads, joint contaminations, and PCR duplications. The sequencing depth was 60× coverage (Supplementary Table S1).

17-mer analysis and estimation of genome size and heterozygosity

Based on the K-mer analysis, 17-mer was selected for analysis, and the *jellyfish* software was used to quickly count 17-mer frequencies (Table 1). *GenomeScope* software was then used to fit the spectrum of 17-mers, and a total of 251.12 G second-generation data were selected from 16 libraries. The *A. lancea* genome has high heterozygosity and repetitive sequence content consistent with a complex genome (Figure 1A). The 17-mer frequency distribution was ∼26.1, and the total K-mer count was 4,159,244,135 bp. The genome size of non-repetitive sequences was estimated to be 291.49 Mb, which was approximately 60.60% of the *A. lancea* genome. The low K-mer frequency indicated the error rate was 0.154%.

GC content and distribution

The assembled contigs were analyzed to obtain GC content and statistical distribution status (Figure 1B). Contigs were mainly distributed in regions where the GC content was between 20 and 50% and concentrated in regions with ∼40% GC content. The average coverage was between 5× and 50×. In this region, two centers of gravity existed in the contig distribution, with an average coverage of approximately 10× and 20×, respectively. Taken together with the K-mer distribution curve, it was speculated that the two centers of gravity corresponded to a heterozygous peak and a homozygous peak, respectively. In the region with the coverage above 50×, there was also a certain amount of contig distribution. It is possible that this was caused by repetitive sequences thus explaining the high heterozygosity rate and repeat sequence ratio of the *A. lancea* genome resulting in difficulties in splicing. According to the GC depth scatter diagram, the GC content of the *A. lancea* genome was estimated to be approximately 38.40%, which was higher than

Bioscience Reports (2020) **40** BSR20202709
https://doi.org/10.1042/BSR20202709

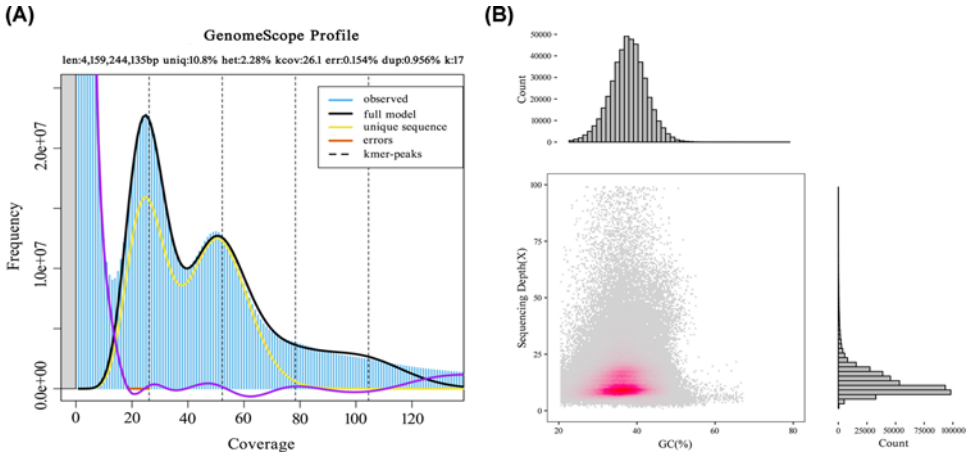


Figure 1. The 17-mer and GC-depth distribution of *A. lancea* genome

(**A**) The distribution curve of 17-mer. (**B**) The GC-depth distribution of *A. lancea* genome.

Table 2 Genomic information statistics of *A. lancea*

	Scaffold		Contig	
	Length (bp)	Number	Length (bp)	Number
max_len	289109		289109	
N10	4256	68307	2857	94469
N20	2477	211142	1622	299948
N30	1629	438525	1071	629298
N40	1120	775151	750	1110774
N50	778	1260281	533	1790925
N60	526	1967071	375	2751303
N70	335	3047918	253	4135094
N80	190	4851250	165	6261804
N90	119	7855566	112	9405484
Total_length	4508659206	4277961587		
number ≥ 100 bp	12088217	13462724		
number ≥ 2000 bp	311077	199719		
GC_rate	0.367		0.384	

that of *Gastrodia elata* (33.4%), *Scutellaria baicalensis* (34.3%) and *Platycodon grandifloras* (36.3%), and lower than that of *Sorghum bicolor* (42.8%) and *Zea mays* (46.8%), but similar to that of *Rosa roxburghii* (38.4%) and *Helianthus annuus* (38.9%). Therefore, the *A. lancea* genome was of mid-GC content.

Results of preliminary assembly of genomic data

The 251.12 Gb clean bases were used for preliminary genome assembly after the sequence data were filtered to obtain the preliminary genome sequence. An important indicator of the quality of genome assembly is the Contig N50 value (Table 2). In the *A. lancea* genome, the longest assembled sequence had a length of 289,109 bp, and the length of the N50 was 533 bp. We obtained 4,277,961,587 scaffolds with a total length of 4,508,659,206 bp after further assembly of contigs into scaffolds using *SOAPdenovo* software. The length of the N50 scaffold was 778 bp, and the GC content was 38.4%.

Genome comparison within the Compositae family

The key word 'Asteraceae' (synonymous with Compositae) was used to search the genome database of National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/genome/) to obtain the genomic information of 11 Compositae family plants, and this information was compared with the genomic data of *A. lancea* (Supplementary Table S2). The published data show that the genomes of 11 species of Compositae plants ranged in size from 121.712

Table 3 SSR types detected in the *A. lancea* sequences

Searching item	Number	Ratio (%)
total_SSR_ number	27582	
total_SSR_ length (bp)	476561	
Number of SSR-containing sequences	25144	91.16%
Number of sequences containing more than one SSR	2212	8.02%
Total number of identified SSR	26034	100.00%
Mononucleotide	14250	54.74%
Dinucleotide	7219	27.73%
Trinucleotide	4217	16.20%
Tetranucleotide	273	1.05%
Pentanucleotide	26	0.10%
Hexanucleotide	49	0.19%

to 4,159.24 Mb. The genome size of *A. lancea* is 4,159.24 Mb and ranks largest among the Compositae plants, followed by *Helianthus annuus* (3,027.84 Mb), *Chrysanthemum seticuspe* (2,721.84 Mb), and *Lactuca sativa* (2,380 Mb). The Compositae plant with the smallest genome is *Silphium perfoliatum* with a genome size of 3,027.84 Mb (Supplementary Figure S2).

If the GC content is higher than 65% or lower than 25%, it may cause sequence bias on the Illumina sequencing platform, and thus seriously compromise genome assembly [23,24]. The GC content of *A. lancea* genome is 38.4%, which is slightly lower than that of Compositae *Helianthus annuus* and higher than that of other plants in the Compositae family, indicating that the GC content of the *A. lancea* genome is in an acceptable range for genome assembly and thus did not affect the quality of genome assembly. Genome heterozygosity rates were compared across the 12 Compositae plants. The heterozygosity rate of *A. lancea* was the highest (2.28%), followed by *Artemisia annua* (1.0-1.5%) indicating that the genome of *A. lancea* is a complex genome with high heterozygosity and a high repeat ratio.

SSR markers

A total of 26,034 SSRs were identified, and the most abundant type of repeat was the mononucleotide, which accounted for 54.74% of the observed SSRs, followed by dinucleotide (27.73%), trinucleotide (16.20%), tetranucleotide (1.05%), pentanucleotide (0.10%), and hexanucleotide (0.19%) repeats (Table 3). The most common repeat type was repeated ten times (6620, 25.43%), and the second most common repeated six times (3193, 12.26%) (Figure 2). Among the dinucleotide motifs, GA/TC was the most abundant (31.17%), followed by AG/CT (26.15%), CA/TG (11.75%), AC/GT (11.04%), TA/TA (10.06%), AT/AT (9.68%), GC/GC (0.08%), and CG/CG (0.07%) repeats. The predominant trinucleotide motifs, TTC/GAA, CCA/TGG, and TGA/TCA repeats accounted for 7.23, 6.90, and 5.81%, respectively (Figure 3).

Gene prediction and annotation

A total of 93,434 unigenes (66% of all unigenes) were annotated (Supplementary Table S3), where the TrEMBL database had the greatest match (92,046 unigenes were annotated, 65%), followed by the NR database (91,165, 64%) and Swiss-Prot database (60,411, 42%). A total of 48,493 putative genes were classified into KOG functional categories. The general function prediction only represented the largest group (10,836; 22.35%), followed by signal transduction mechanisms (4,603; 9.50%) and post-translational modification, protein turnover, and chaperones (4,579; 9.44%) (Supplementary Figure S3).

Based on sequence homology, the 33,896 assembled transcripts were assigned GO terms and further classified into the categories of biological process, cellular component, and molecular function (Figure 4). Among the biological processes, the most represented category was metabolic processes (45.87%), followed by cellular processes (23.98%), and localization processes (17.59%). When sorted based on cellular component, the membrane was the most represented category (30.79%), followed by intracellular (21.42%) and protein-containing complexes (20.35%). For molecular function, the two most represented categories were binding (51.84%) and catalytic activity (38.28%).

There were 34,929 putative genes assigned to 42 KEGG pathways. A total of 20,862 genes were associated with 12 metabolic pathways; 4,948 (23.72%) were involved in carbohydrate metabolism, followed by global and overview maps (3,025; 14.5%), amino acid metabolism (3,015; 14.5%), and lipid metabolism (2,181; 10.5%). A total of 12,486 genes were associated with organismal systems pathways, 8,749 genes with environmental information processing, 8,336 genes with genetic information processing, 6,654 genes with cellular processes, 5,624 genes with viral diseases

Bioscience Reports (2020) **40** BSR20202709
https://doi.org/10.1042/BSR20202709

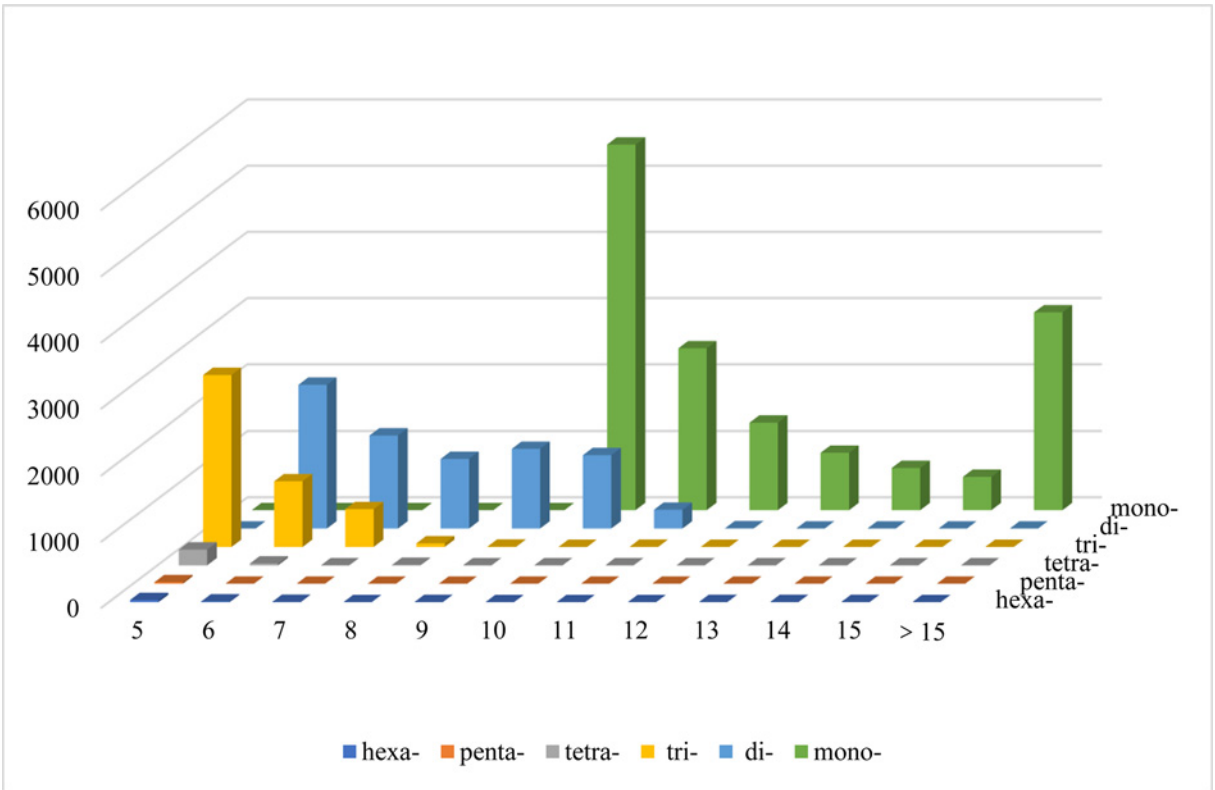


Figure 2. Distribution of various classes of simple repeat motifs with different numbers of repeats in the *A. lancea* genome

X-axis, number of SSR repeats; Y-axis, frequency of SSR type.

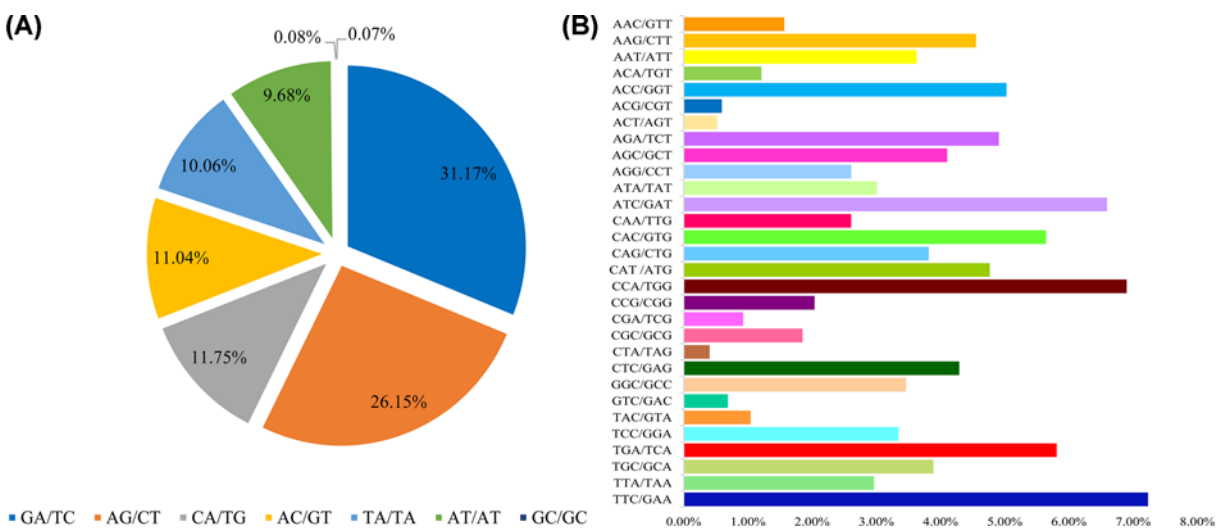


Figure 3. Percentage of different motifs in dinucleotide and trinucleotide repeats in *A. lancea*

(**A**) Frequency of different dinucleotide SSR motifs. (**B**) Frequency of different trinucleotide SSR motifs.

and pathways, 4,957 genes with human diseases, and 675 genes with other specific processes and pathways (Figure 5).

The 91,165 unigenes with NR annotation were compared with other species (Supplementary Figure S4). The transcriptome of *Cynara cardunculus* had the largest number of genes similar to the *A. lancea* transcriptome, accounting for 50.78%, followed by the *Lactuca sativa* (9.96%), *Homo sapiens* (9.88%), and *Rhizoctonia solani* (8.39%).

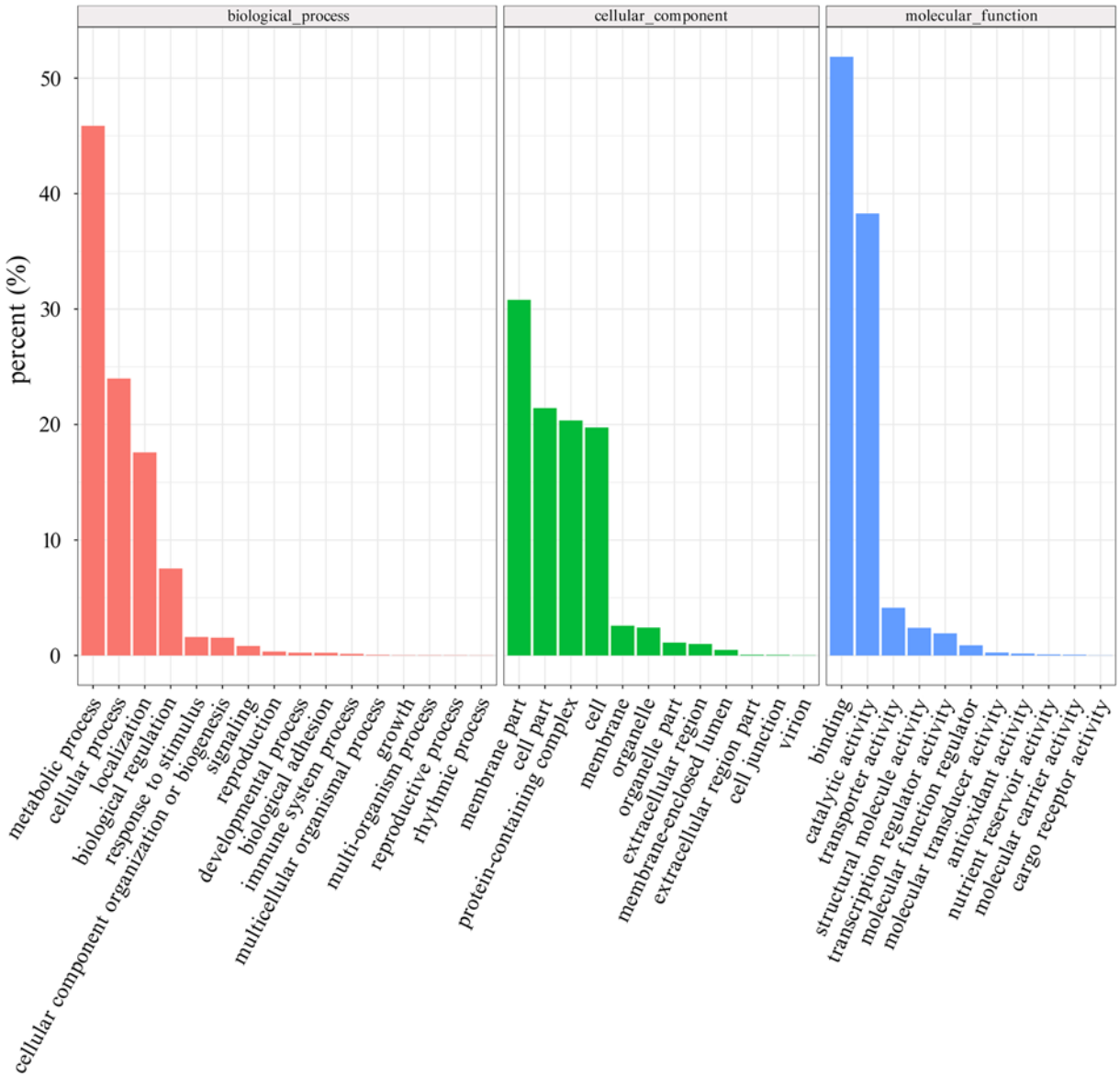


Figure 4. GO functional classification of *A. lancea* unigenes

Discussion

Asteraceae (Compositae) is the largest family of angiosperms in terms of numbers of species, which include many species of significant medicinal and ornamental value. Its special inflorescence composition, corolla type, and other morphological characteristics have important taxonomic significance [25]. *A. lancea* can be divided into two types, namely the Maoshan and Dabieshan types [26,27]. The Maoshan type is distributed primarily in the Maoshan area of Jiangsu Province. However, in recent years, the wild *A. lancea* in Jiangsu province has been gradually becoming endangered and has been listed as one of the four endangered medicinal plants in Jiangsu province [28]. Therefore, the study of the *A. lancea* genome will have important ecological significance.

In the current study, the whole genome of *A. lancea* was sequenced for the first time. Sequencing and analysis of the genome was performed using the NGS method, and the genome size was automatically estimated to be 4.16 Gb using *GenomScope* software. The GC content of the *A. lancea* genome was 38.4%, and the heterozygosity rate and repeat sequence ratios were 2.28 and 89.2%, respectively. The high heterozygosity rate and repeat sequence content indicates a complex genome. To obtain a high-quality whole genome map of *A. lancea*, a strategy combining the PacBio and Illumina sequencing platforms, supplemented by the High-throughput Chromosome Conformation Capture (Hi-C)

Bioscience Reports (2020) **40** BSR20202709
https://doi.org/10.1042/BSR20202709

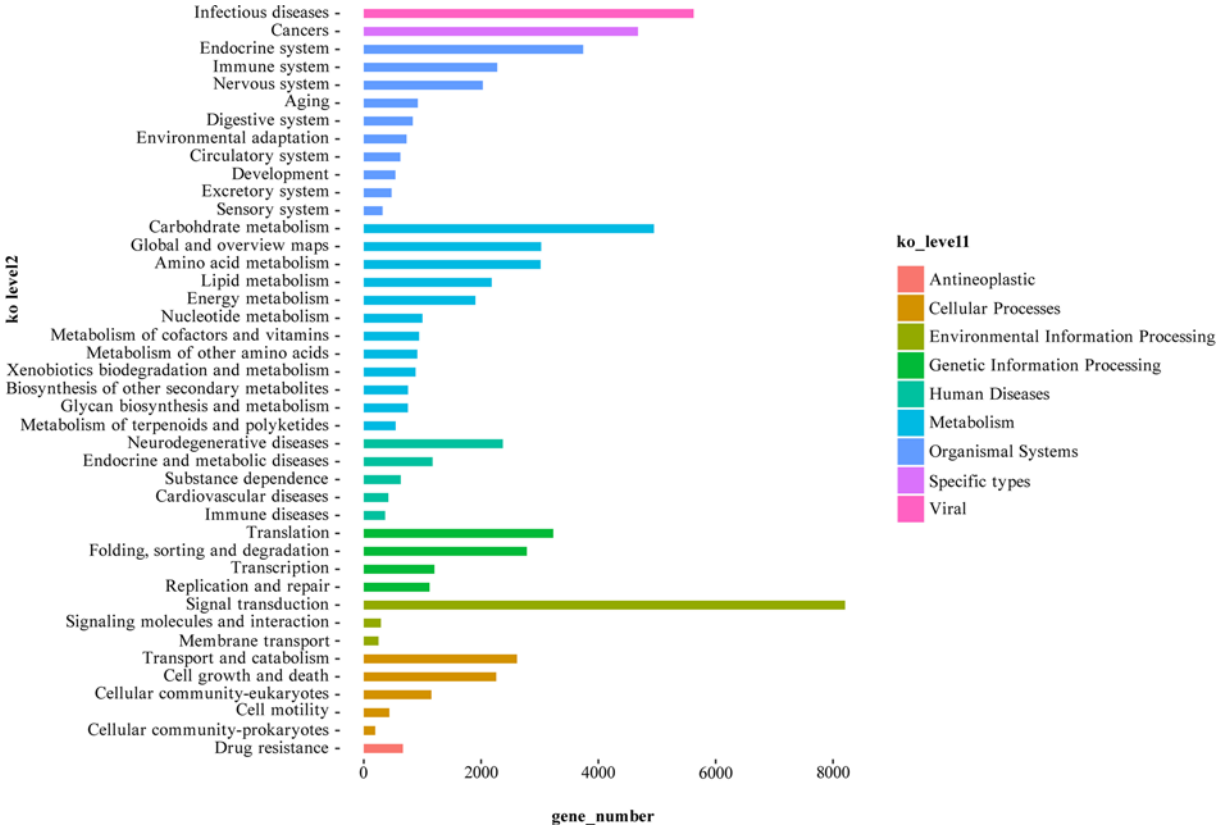


Figure 5. KEGG functional classification of *A. lancea* unigenes

technique is recommended. Genome size, also known as C-value or Constant-value, is an important genetic feature of an organism [29]. The genome 2C-value of *A. lancea* was estimated previously by Deng et al. to be 9.73 pg (1 pg = 978 Mbp) by flow cytometry [8], which was close to the results of our genome survey. There are great differences in genome size among species within the Compositae family. Vinogradov [30] showed that the genomes of threatened plants (whose populations are now on the decline) may actually be larger than expected compared with those of their relatives. The genome of *A. lancea* ranks first in the Compositae family in regard to size at 4,159.24 Mb. This shows that *A. lancea* is more endangered than other plants of the Compositae family.

The genomic SSR analysis of *A. lancea* showed that there were great differences in the SSR content. The SSR types of *A. lancea* were abundant and with high frequencies of mononucleotide, dinucleotide, and trinucleotide repeats. This is similar to plants such as *Dioscorea zingiberensis* [31], *Rosa roxburghii* [32], *Dracaena cambodiana* [33], and *Apocynum venetum* [34], but not consistent with the majority of plants where dinucleotide and trinucleotide repeats are the main types of SSR [35–37]. Among the dinucleotide repeat motifs, the GA/TC and AG/CT repeats were the two most abundant types, which accounted for 31.17 and 26.15%, respectively, followed by CA/TG (11.75%). Among the trinucleotide repeat motifs, the TTC/GAA was most abundant, which accounted for 7.23%, followed by CCA/TGG (6.90%) and TGA/TCA (5.81%). The SSRs of *A. lancea* are polymorphic, thus laying a foundation for using these SSR molecular markers in screening for genetic diversity.

A total of 93,434 *A. lancea* genes, 66% of the *A. lancea* genome, matched known genes in common databases. It is speculated that the reason for the remaining uncommented genes may be due to either non-coding or incomplete sequences [38]. According to the NR annotation, *Cynara cardunculus* genes were the most commonly annotated with *A. lancea* indicating that *A. lancea* has the closest relationship with *Cynara cardunculus*. Comparing the functional analyses of the *A. lancea* genome using the KOG, KEGG, and GO classification systems, it was found that the number of transcripts annotated to KOG was the largest, with the most abundant genes related to general function, signal transduction mechanism and post-translational modification, protein turnover, and chaperones. The results of the present study will play an important role in future whole-genome sequencing projects and provide a rich resource for

future functional studies, which will greatly enhance our understanding of the genetic regulatory mechanisms of *A. lancea*.

Competing Interests

The authors declare that there are no competing interests associated with the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China [grant numbers 82073957, 81703633, 81773853, 81973432, 81803675]; the Anhui Provincial Natural Science Foundation [grant number 1808085QH290]; the Special Fund for Guiding Local Science and Technology Development, awarded by the Central Government of Anhui Province [grant number YDZX20183400004233]; the key project of natural science research in Universities of Anhui Province [grant number KJ2019A0461); the Key Project at the Central Government Level: The Ability Establishment of Sustainable Use for Valuable Chinese Medicine Resources [grant number 2060302]; and the CAMS Innovation Fund for Medical Sciences [grant number 2019-I2M-5-065].

Author Contribution

Tingyu Shan and Junxian Wu contributed significantly to data analyses and the writing of the manuscript. Daqing Yu and Jin Xie contributed reagents and analysis tools. Qingying Fang contributed the samples and data. Liangping Zha and Huasheng Peng designed and supervised the study and provided critical revision of the manuscript. All authors read and approved the final manuscript.

Abbreviations

GO, Gene Ontology database; KEGG , Kyoto Encyclopedia of Genes and Genomes; KOG , euKaryotic Clusters of Orthologous Groups; NGS, next-generation sequencing; NR, non-redundant protein database; SSR, simple sequence repeat.

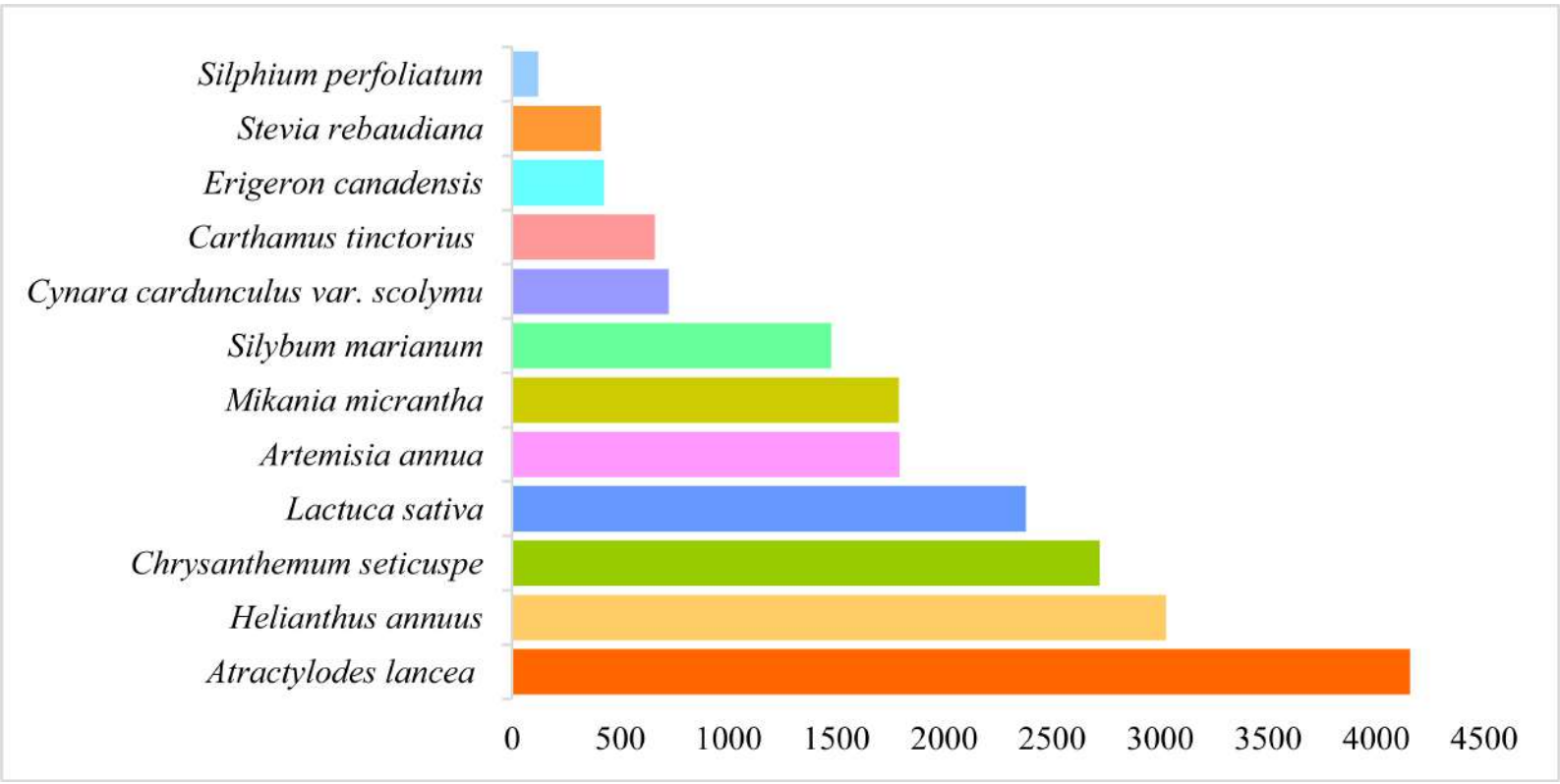
References

1 Liu, Q.T., Zhang, S.S., Yang, X.H. et al. (2016) Differentiation of essential oils in *Atractylodes lancea* and *Atractylodes koreana* by gas chromatography with mass spectrometry. *J. Sep. Sci* **39**, 4773–4780, https://doi.org/10.1002/jssc.201600574
2 Zhang, L., Ou, Y.Z., Zhao, M. et al. (2010) Simultaneous determination of atractylone, hinesol, β-eudesmol, atrctylodin in *Atractylodes lancea* and hierarchical cluster analysis. *China J. Chin. Mater. Med* **35**, 725–728
3 Koonrungsesomboon, N., Na-Bangchang, K. and Karbwang, J. (2014) Therapeutic potential and pharmacological activities of *Atractylodes lancea* (Thunb.) DC. *Asian Pac. J. Trop. Med.* **6**, 421–428, https://doi.org/10.1016/S1995-7645(14)60069-9
4 Deng, A.P., Li, Y., Wu, Z.T. et al. (2016) Advances in studies on chemical compositions of *Atractylodes lancea* and their biological activities. *China J. Chin. Mater. Med.* **41**, 3904–3913
5 Xie, J., Peng, F., lei, Yu and Peng, C. (2018) Pharmacological efects of medicinal components of *Atractylodes lancea* (Thunb.) DC. *Chin. Med.* **13**, 59–69
6 Shi, S.J., Qin, Z., Kong, S.Z. et al. (2012) Effective component selection of atraotydin extract against influenza virus. *Shih-chen Kuo I Kuo Yao* **23**, 565–566
7 Qin, J., Wang, H.Y., Zhuang, D. et al. (2019) Structural characterization and immunoregulatory activity of two polysaccharides from the rhizomes of *Atractylodes lancea* (Thunb.) DC. *Int. J. Biol. Macromol.* **136**, 341–351, https://doi.org/10.1016/j.ijbiomac.2019.06.088
8 Deng, J., Ni, J.J., Wan, W.T. et al. (2015) Nuclear genome size and chromosome analysis in *Atractylodes lancea*. *Lishizhen Med. Mater. Med. Res* **26**, 2499–2501
9 Zhang, D.C., Li, D.L. and Gao, X.Q. (1992) A karyotype study of *Atractylodes lancea* (Thunb.) DC. *J. Anhui Norm. Univer.* **4**, 71–75
10 Li, X.L., Wang, X.S., Cheng, S.H. et al. (2015) Chromosome karyotype analysis of six populations in *Atractylodes* DC. *J. Plant Genet. Resour.* **16**, 185–191
11 Deng, J., Wan, Q.Y., Gong, L. et al. (2017) Cloning and analysis of DXS gene from *Atractylodes lancea*. *Chin. J. Exp. Tradit. Med. Formulae* **23**, 39–44
12 Chen, L.N., Wan, Q.Y., Deng, J. et al. (2019) Cloning and analysis of two DXR genes (AlDXR) in *Atractylodes lancea*. *Mol. Plant Breed.* **17**, 4249–4256
13 Liu, Q., Cao, X.Y., Jing, J.H. et al. (2007) Cloning and analysis of HMGR gene conserved fragments in *Atractylodes lancea*. *Chin. Tradit. Herb Drugs* **38**, 1551–1554
14 Jiang, L., Gu, W., Chao, J.G. et al. (2017) Gene cloning of farnesyl pyrophosphate synthase in *Atractylodes lancea* and its expression pattern analysis. *Chin. Tradit. Herb Drugs* **48**, 760–766
15 Bowers, J.E., Pearl, S.A. and Burke, J.M. (2016) Genetic mapping of millions of SNPs in Safflower (*Carthamus tinctorius* L.) via whole-genome resequencing. *G3-Genes Genom. Genet.* **6**, 2203–2211
16 Shen, Q., Zhang, L., Liao, Z. et al. (2018) The genome of *Artemisia annua* provides insight into the evolution of Asteraceae family and Artemisinin biosynthesis. *Mol. Plant* **11**, 776–788, https://doi.org/10.1016/j.molp.2018.03.015
17 Metzker, M.L. (2010) Sequencing technologies-the next generation. *Nat. Rev. Genet* **11**, 31–46, https://doi.org/10.1038/nrg2626
18 Kumar, K.R., Cowley, M.J. and Davis, R.L. (2019) Next-generation sequencing and emerging technologies. *Semin. Thromb. Hemost.* **45**, 661–673

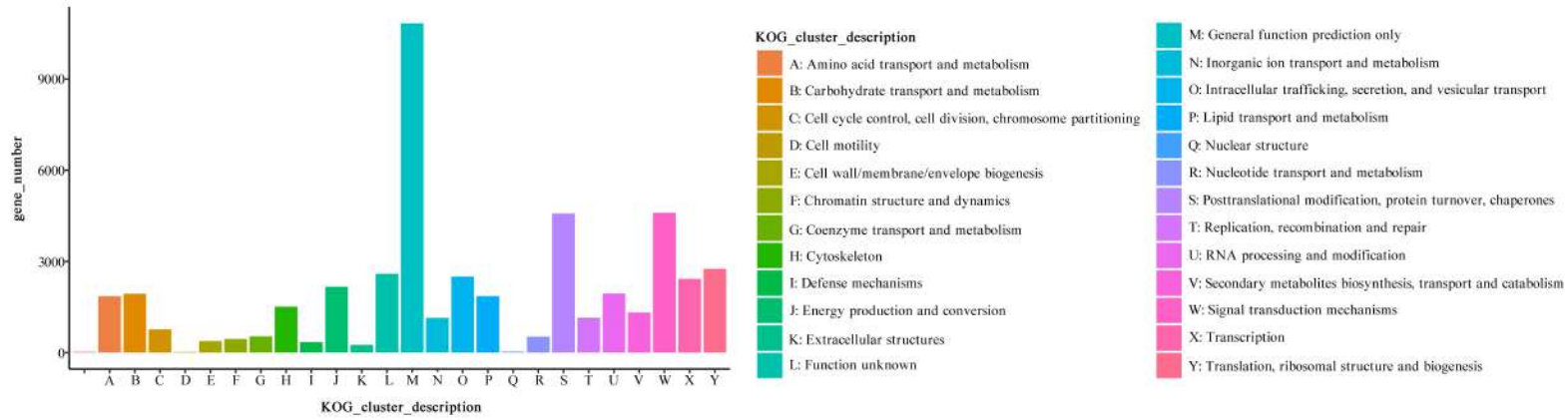
19 Bi, Q.X., Zhao, Y., Cui, Y.F. et al. (2019) Genome survey sequencing and genetic background characterization of yellow horn based on next-generation sequencing. *Mol. Biol. Rep.* **46**, 4303–4312, https://doi.org/10.1007/s11033-019-04884-7
20 Wang, C.R., Yan, H.D., Li, J. et al. (2018) Genome survey sequencing of purple elephant grass (*Pennisetum purpureum* Schum 'Zise') and identification of its SSR markers. *Mol. Breed.* **38**, 94–103, https://doi.org/10.1007/s11032-018-0849-3
21 Ziya Motalebipour, E., Kafkas, S., Khodaeiaminjan, M. et al. (2016) Genome survey of pistachio (*Pistacia vera* L.) by next generation sequencing: Development of novel SSR markers and genetic diversity in Pistacia species. *BMC Genomics* **17**, 998–1012, https://doi.org/10.1186/s12864-016-3359-x
22 Murray, M.G. and Thompson, W.F. (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8**, 4321–4325, https://doi.org/10.1093/nar/8.19.4321
23 Wu, Y.F., Xiao, F.M., Xu, H.N. et al. (2014) Genome survey in *Cinnamomum camphora* L. *Plant Genet. Resour.* **15**, 149–152
24 Aird, D., Ross, M.G., Chen, W.S. et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, 1–14, https://doi.org/10.1186/gb-2011-12-2-r18
25 Zhang, D.E. and Zheng, H.C. (2000) *Conservation of Endangered Medicinal Wildlife Resources in China*, Sencond Military Medical University Press, Shanghai
26 Yu, D.Q., Han, X.J., Shan, T.Y. et al. (2019) Microscopic characteristic and chemical composition analysis of three medicinal plants and surface frosts. *Molecules* **24**, 4548–4546, https://doi.org/10.3390/molecules24244548
27 Wang, F., OuYang, Z., Guo, L.P. et al. (2014) Comprehensive chemical pattern recognition of Atractylodis Rhizoma. *China J. Chin. Mater. Med.* **39**, 2536–2541
28 Liu, C.S. (2016) *Pharmaceutical Botany*, China Press of Traditional Chinese Medicine, Beijing
29 Dolezel, J., Bartos, J., Voglmayr, H. et al. (2003) Nuclear DNA content and genome size of trout and human. *Cytom. Part A* **51**, 127–128
30 Vinogradov, A.E. (2003) Selfish DNA is maladaptive: evidence from the plant Red List. *Trends Genet.* **19**, 609–614, https://doi.org/10.1016/j.tig.2003.09.010
31 Zhou, W., Li, B., Li, L. et al. (2018) Genome survey sequencing of *Dioscorea zingiberensis*. *Genome* **61**, 567–574, https://doi.org/10.1139/gen-2018-0011
32 Lu, M., An, H. and Li, L. (2016) Genome survey sequencing for the characterization of the genetic background of *Rosa roxburghii* Tratt and leaf ascorbate metabolism genes. *PLoS ONE* **11**, e0147530, https://doi.org/10.1371/journal.pone.0147530
33 Ding, X., Mei, W., Huang, S. et al. (2018) Genome survey sequencing for the characterization of genetic background of *Dracaena cambodiana* and its defense response during dragon's blood formation. *PLoS ONE* **13**, e0209258, https://doi.org/10.1371/journal.pone.0209258
34 Li, G.Q., Song, L.X., Jin, C.Q. et al. (2019) Genome survey and SSR analysis of *Apocynum venetum*. *Biosci. Rep.* **39**, https://doi.org/10.1042/BSR20190146
35 Kumpatla, S.P. and Mukhopadhyay, S. (2005) Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome* **48**, 985–998, https://doi.org/10.1139/g05-060
36 Hou, R., Bao, Z., Wang, S., Su, H., Li, Y., Du, H. et al. (2011) Transcriptome sequencing and de novo analysis for Yesso scallop (*Patinopecten yessoensis*) using 454 GS FLX. *PLoS ONE* **6**, e21560, https://doi.org/10.1371/journal.pone.0021560
37 Zhou, X., Dong, Y., Zhao, J., Huang, L., Ren, X., Chen, Y. et al. (2016) Genomic survey sequencing for development and validation of single-locus SSR markers in peanut (*Arachis hypogaea* L.). *BMC Genomics* **17**, 420–434, https://doi.org/10.1186/s12864-016-2743-x
38 Hou, R., Bao, Z., Wang, S., Su, H., Li, Y., Du, H. et al. (2011) Transcriptome sequencing and de novo analysis for Yesso scallop (*Patinopecten yessoensis*) using 454 GS FLX. *PLoS ONE* **6**, e21560, https://doi.org/10.1371/journal.pone.0021560


1 cm

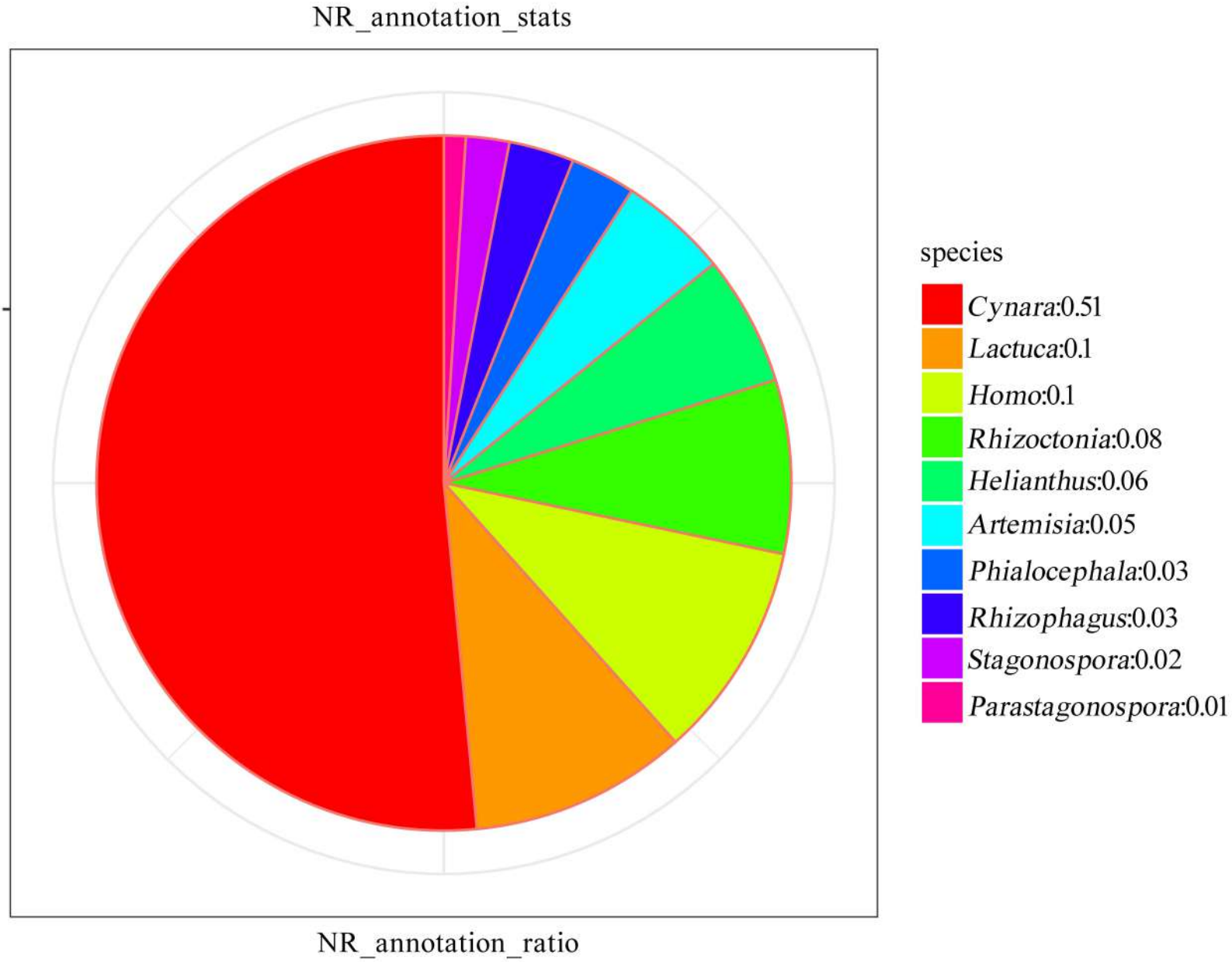
Supplementary Figure S1. Plants and rhizome of *A.lancea*


Silphium perfoliatum
Stevia rebaudiana
Erigeron canadensis
Carthamus tinctorius
Cynara cardunculus var. scolymu
Silybum marianum
Mikania micrantha
Artemisia annua
Lactuca sativa
Chrysanthemum seticuspe
Helianthus annuus
Atractylodes lancea
0
500
1000
1500
2000
2500
3000
3500
4000
4500

Supplementary Figure S2. Comparison the genome size of 12 species of Compositae


gene_number
9000
6000
3000
0
A B C D E F G H I J K L M N O P Q R S T U V W X Y
KOG_cluster_description
KOG_cluster_description
A: Amino acid transport and metabolism
B: Carbohydrate transport and metabolism
C: Cell cycle control, cell division, chromosome partitioning
D: Cell motility
E: Cell wall/membrane/envelope biogenesis
F: Chromatin structure and dynamics
G: Coenzyme transport and metabolism
H: Cytoskeleton
I: Defense mechanisms
J: Energy production and conversion
K: Extracellular structures
L: Function unknown
M: General function prediction only
N: Inorganic ion transport and metabolism
O: Intracellular trafficking, secretion, and vesicular transport
P: Lipid transport and metabolism
Q: Nuclear structure
R: Nucleotide transport and metabolism
S: Posttranslational modification, protein turnover, chaperones
T: Replication, recombination and repair
U: RNA processing and modification
V: Secondary metabolites biosynthesis, transport and catabolism
W: Signal transduction mechanisms
X: Transcription
Y: Translation, ribosomal structure and biogenesis

Supplementary Figure S3.KOG functional classification of *A. lancea*unigenes.


NR_annotation_stats
species
Cynara:0.51
Lactuca:0.1
Homo:0.1
Rhizoctonia:0.08
Helianthus:0.06
Artemisia:0.05
Phialocephala:0.03
Rhizophagus:0.03
Stagonospora:0.02
Parastagonospora:0.01
NR_annotation_ratio

Supplementary Figure S4. NR annotated homologous species distributions

Supplementary Table S1. Statistics of filtering data

Lib id	Length	Raw data	Clean data	GC rate
190626_I9_V300022484_L3_RHIshvDAAAA-573	150	14,518,852,200	14,249,431,800	38.98
190626_I9_V300022484_L3_RHIshvDAAAA-574	150	16,785,220,500	16,471,137,000	38.91
190626_I9_V300022484_L3_RHIshvDAAAA-575	150	14,729,106,000	14,448,511,200	38.9
190626_I9_V300022484_L3_RHIshvDAAAA-576	150	13,615,597,500	13,348,510,200	38.93
190626_I9_V300022484_L3_RHIshvDAAAA-577	150	15,861,435,600	15,566,351,400	38.89
190626_I9_V300022484_L3_RHIshvDAAAA-578	150	16,151,568,000	15,851,755,800	38.94
190626_I9_V300022484_L3_RHIshvDAAAA-579	150	17,966,042,100	17,646,709,500	38.87
190626_I9_V300022484_L3_RHIshvDAAAA-580	150	14,990,523,000	14,716,546,800	38.94
190626_I9_V300022484_L4_RHIshvDAAAA-573	150	15,333,206,400	14,993,685,600	38.88
190626_I9_V300022484_L4_RHIshvDAAAA-574	150	17,745,398,100	17,352,470,700	38.8
190626_I9_V300022484_L4_RHIshvDAAAA-575	150	15,459,255,300	15,108,670,500	38.81
190626_I9_V300022484_L4_RHIshvDAAAA-576	150	14,491,109,700	14,161,733,400	38.8
190626_I9_V300022484_L4_RHIshvDAAAA-577	150	16,822,270,200	16,451,535,000	38.78
190626_I9_V300022484_L4_RHIshvDAAAA-578	150	17,067,861,600	16,689,940,500	38.84
190626_I9_V300022484_L4_RHIshvDAAAA-579	150	18,935,608,800	18,532,550,100	38.77
190626_I9_V300022484_L4_RHIshvDAAAA-580	150	15,874,163,700	15,529,107,600	38.84

Supplementary Table S2. Genome assembly statistics of *A. lancea* and comparisons to 11 genomes of Compositae

Species	Gene size (Mb)	Number of protein-coding genes	GC content (%)	Hetero zygosity (%)	Repeat (%)	References
Atractylodes lancea	4159.24	n.c	38.4	2.28	89.2	This study
Carthamus tinctorius (safflower)	661.938	n.c	37.3	n.c	n.c	Bower, J.E, 2016
Stevia rebaudiana	411.38	n.c	34.6	n.c	n.c	n.c
Chrysanthemum seticuspe	2721.84	n.c	36.1	n.c	72.5	Hideki, 2019
Silphium perfoliatum	121.712	n.c	36.7	n.c	n.c	n.c
Silybum marianum	1477.58	n.c	37.2	n.c	n.c	n.c
Erigeron canadensis	425.611	n.c	34	0.203	42.9	Laforest et al., 2020 Peng et al., 2014
Mikania micrantha	1790.64	46329	36.2	n.c	n.c	n.c
Cynara cardunculus var. *scolymus*	725.198	38406	36.75	n.c	n.c	Caglione et al., 2016
Lactuca sativa	2380	38919	37.75	n.c	n.c	Verwaaijen et al., 2017 Reye et al., 2016
Artemisia annua	1792.86	66918	34.1	1.0-1.5	60.1	Shen et al., 2018
Helianthus annuus	3027.84	73839	38.91	n.c	n.c	Bock et al., 2014 Badouin et al., 2017

Supplementary Table S3. Statistics of gene functional annotation

Data_base	annotated_number	annotated_ratio
GO	33896	24%
KEGG	34929	24%
KOG	48493	34%
nr	91165	64%
pfam	58121	41%
swiss_prot	60411	42%
TrEMBL	92046	65%
Total	93434	66%