High-Quality Chromosome-Level De Novo Assembly of the Trifolium repens

Hongjie Wang
College of Grassland Science, Qingdao Agricultural University
Yongqiang Wu
College of Grassland Science, Qingdao Agricultural University
Yong He
College of Grassland Science, Qingdao Agricultural University
Guoyu Li
College of Grassland Science, Qingdao Agricultural University
Lichao Ma
Key Laboratory of National Forestry and Grassland Administration on Grassland Resources and Ecology in the Yellow River Delta
Shuo Li
College of Grassland Science, Qingdao Agricultural University
Jianwei Huang
Berry Genomics Corporation
Guofeng Yang (yanggf@qau.edu.cn)
College of Grassland Science, Qingdao Agricultural University

Research Article

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

Posted Date: March 21st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2631739/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Genomics on June 13th, 2023. See the published version at https://doi.org/10.1186/s12864-023-09437-8.

Abstract

Background: White clover (*Trifolium repens L.*), an excellent perennial legume forage, is a heterotetraploid native to southeastern Europe and southern Asia. It has high feeding, 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 white clover 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 were used to generate a 1096 Mb genome size of *T. repens*, with contigs of N50 = 14 Mb, scaffolds of N50 = 65 Mb, and BUSCOs 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 most 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 for white clover obtained at the chromosomal level using PacBio third-generation Hi-Fi 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.

Background

White clover (*Trifolium repens L.*), 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 feeding, 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 to its soil moisturization effect. White clover exhibits excellent growth when mixed with forages of the family Gramineae, forming a natural grassland that can effectively prevent swelling disease in livestock[6]. The forage also improves the feed value of grassland, thus playing an important role in the stable development of the grassland ecosystem. Furthermore, 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 technologies to generate a high-quality chromosome-level genome assembly of white clover[15, 16]. Compared with the previously published 841 Mb (N50 = 122 kb) genome assembly of white clover, the assembly generated in this study contains 202 scaffolds (~ 1096 Mb) spanning N50 = 65 Mb and has a significantly improved quality[17]. 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[18]. 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. 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[19]. 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[20, 21].

Results

Genome-Survey, Sequencing, and Assembly

This study evaluated the size, repeatability, heterozygosity, and other genome parameters of the white clover. After quality control, Illumina sequencing yielded 59 Gb of data[18]. Blasting 10,000 randomly selected clean reads against the NT library revealed a 98.79% mapping. Moreover, K-mer analysis performed to estimate the complexity of the genome further predicted genome size of 1075 Mb, with 1.68% repeat and 68.80% heterozygous sequences[22].

Third-generation Hi-Fi sequencing (TGS) technology developed by PacBio was used to initially assemble the white clover genome based on traditional next-generation sequencing (NGS) data assembly methods[23]. 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 splicing using PacBio Hi-Fi an effective assembly strategy[24]. 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 1.63kbp.

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), closer to the previously assembled *Trifolium pratense* genome (33.60%)[25, 26]. To evaluate the quality and integrity of the assembly, we compared the sequencing data

with the assembly results and found that the properly paired mapping was 89.22%, while BUSCOs assembly assessment integrity was 98.50%. These results indicate that the assembly results had good integrity[27].

Table 1
Summary statistic for the *Trifolium repens* genome

	Assembly	
Genome assembly	Estimated genome size	1075Mb
	Total length of assembly	1096Mb
	Number of contigs	380
	Contig N50	14Mb
	Largest contig	53Mb
	Number of scafolds	202
	Scafold N50	65Mb
	Chromosome coverage(%)	95.06%
	GC content of genome	33.64%
	Annotation	
		Total length
Transposable elements	Total	672Mb(61.37%)
	Retrotransposon	448Mb(40.91%)
	DNA Transposon	140Mb(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	3,604bp
	Mean coding sequence length	1,592bp

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	29006	1430	7.6	293	449
Glycine max	54881	1391	8	295	413
Trifolium medium	119102	306	1.4	219	172
Cicer arietinum	28772	1393	7.7	291	418
Medicago truncatula	36079	1428	6.9	324	393
Trifolium repens	90128	1592	5	341	490

Scaffold Construction and Curation

The high-throughput chromatin conformation capture (Hi-C) technology utilizes the entire cell nucleus to fix and capture the mutual chromosomal sites[19, 28]. 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[12, 19, 28]. 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 better[17]. Moreover, after Hi-C-assisted assembly, 1.04 Gb of genome sequences were identified on 16 chromosomes, accounting for 95.06% of the entire genome. After Hi-C-assisted assembly, it was observed that the genetic material exchange was much stronger within than between chromosomes[29]. The heat map showing the genome interaction of the Hi-C-assisted assembly further verified the accuracy of the assembly results (Fig. 1). 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, and additional 4092 simple repeats were also found in the assembled genome. We also predicted 13 types of ncRNA, totaling 15520 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[30]. We annotated 88,094, 61,830, 77,722, 52,992, and 26,979 genes using NR, Swiss-Prot, 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 S1). 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 eight of *M. truncatula* had a good collinear relationship (Fig. 3), indicating their chromosomal conservation after species divergence[26].

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 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)[31]. Furthermore, the 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 could explain its excellent agronomic traits (Table S2).

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[32]. White clover was most closely related 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[33]. 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 the 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, in this study, a high-quality tetraploid white clover genome was assembled 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. The results revealed a 1096 Mb genome size of *T. repens*, with contig N50 = 14 Mbp, scaffold N50 = 65 Mbp, and BUSCOs = 98.50%. The newly assembled white clover genome had better continuity and integrity than its previously reported reference genome[17]. Additionally, the assembled genome had higher coverage (95.06%) at the chromosomal level after high-throughput sequencing and Hi-C scaffolding. We also annotated 90,128 high-confidence gene models from the newly assembled genome. 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[34].

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 relationship. 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*[35]. Thus, these species share the same ancestry with *T. repens*. 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[23, 35]. 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.

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 index; 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 was conducted in compliance with the ethical norms of Chinese and international regulations.

DNA isolation and sequencing

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). Thereafter, genome sequencing was performed by Berry Hekang (Beijing, China) using the third-generation PacBio Sequel II assembly sequencing platform. Quality and quantity control of the DNA samples were conducted, 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, PCR amplification. The libraries were then subjected to paired-end (PE) sequencing using Illumina NovaSeq[36–38]. After filtering the reads, 10000 clean reads were randomly selected and blasted against the NCBI non-redundant nucleotide database (NT library) to check for possible external contamination[39]. Subsequently, K-mer analysis was used to estimate the genome size, sample heterozygosity, and genome repeat sequence ratio[40]. 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

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[41]. The library size was then determined by pulsed-field electrophoresis, and the acquired data were filtered and loaded onto smrtlink for CCS (Circular Consensus Sequencing) processing. Thereafter, purge_dups software was used to remove heterozygous contigs[42], while BWA was used to remove pseudo contigs from the genome. The results of genome assembly were

evaluated based on the proportion of matched read pairs and the distribution of inserted fragments. Tblastn, Augustus, and Hmmer tools were used to evaluate the integrity of the single-copy lineologous gene bank[27].

HiC 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. Glycine was then added to the mixture to terminate the chromatin cross-linking reaction, and the treated tissues were collected and frozen in liquid nitrogen. The tissues were then 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[43, 44]. 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[18]. Thereafter, 10,000 pairs of sequencing reads were randomly selected from the Hi-C sequencing library data and blasted against the NT database. The top 10 matched species were sequenced and evaluated to determine whether there was bacterial contamination. The JUICER software was then employed to compare the Hi-C data with the draft genome[45, 46]. We analyzed the Hi-C library results via 3D-DNA comparison to obtain valid Hi-C data and generate the chromosome-level scaffold of the white clover genome[28, 46]. 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.

Genome Annotation

Repetitive sequences of the white clover genome were annotated using homology-based and *ab initio* search methods. The sequences were analyzed and predicted using RepeatMasker, MITE Hunter, LTRharvest, LTR Finder, LTR retriever, RepeatModeler, and MITES. Meanwhile, the LTR transposable elements were identified using structure-based prediction methods[47, 48].

We used MITEs to search the genome for Class II transposition factor mites and involuntary transposition factors less than 2kb in length. The software parameters for LTRharvest and LTR Finder were: -similar 90; -vic 10; -seed 20; -seqids yes -minlenltr 100; -maxlenltr 7000; -mintsd 4; -maxtsd 6; -motif TGCA -motifmis And -D 15000 -d 1000 -L 7000 -l 100 -p 20 -C -M 0.9[49, 50]. Moreover, the parameters settings for RepeatModeler for the *de novo* identification of the repeat sequences in masked genomes were -engine ncbi -pa 60. The RepeatMasker parameters for masking repeat sequences in the genome were -s -nolow -norna -gff -engine ncbi -parallel 20. The tRNA *ab initio* rRNA was predicted using tRNAscan-SE software, and the other types of ncRNA were searched using the Rfam database[51–53]. 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*, *T. medium, V. radiata*, *C. arietinum*, and *G. max* were downloaded. These coding sequences were then subjected to blast (v. 2.2.20) searches against the white clover

genome, and the homologs containing premature stop codons and frameshifts were discarded. White clover RNA-seq data were aligned to the generated contigs using GeMoMa-1.6.1, and a comprehensive transcriptome database was built using PASA (v. 2.0.1)[30]. PASA (v. 2.0.1) was also used to predict the open reading frames, and the resulting database was used to train parameters for the four *de novo* gene prediction software packages. These packages included AUGUSTUS (v. 3.2.2), GeneMarker-ET (v. 4.57), GlimmerHMM (v. 3.0.2), and SNAP[54, 55]. The predictions obtained using these packages were combined using EVM, after which 36,511 genes were retrieved and functionally annotated by blast searches against NR, Swiss-Pro, eggNOG, GO, and KEGG databases. Venn analysis of these databases was then performed to obtain more accurate gene functional annotation information[56].

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 suffix tree data structure[57, 58]. To determine the similarity between sequences, we used OrthoMCL clustering analysis to perform all-VS-All BLAST alignment on gene protein-coding sequences of all selected species (e-value = 1e-5 by default)[59]. 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[59, 60].

The Mafft software was subsequently used for multiple sequence comparisons of supergenes[61]. A suitable base substitution model was selected, followed by constructing a species-based maximum likelihood (ML) phylogenetic tree whose differentiation time was estimated using the RAxML software[32, 62, 63]. Moreover, the mcmctree tool of the PAML 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[64, 65]. The gene families of each species were then analyzed using the Cafe software. The numbers of gene family contractions and expansions on each evolutionary branch were obtained, and their occurrences were assessed. Furthermore, protein-coding sequences were identified using the positive selection approach by distinguishing between synonymous substitutions (Ks) and non-synonymous substitutions (Ka)[66].

The number of Ks in each synonymous locus of the constructed genome was used to detect WGD events[67]. Moreover, Blastp was used to compare the longest protein sequence encoded by the white clover genes. The MCScanX software was subsequently used to filter the comparison results, and the Yn00 tool of the PAML software package was used to calculate the synonymous replacement rate[68]. 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[31, 69]. 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-through chromosome conformation capture
MITEs
miniature inverted repeat transposable elements
LTR
Long terminal repeat
LTR-RT
Long terminal repeat retrotransposons
ncRNA
Non-coding RNA
NR
NCBI nucleotide sequences
GO
Gene Ontology
KEGG
Kyoto Encyclopedia of Genes and Genomes
WGD
Whole Genome Duplications.

Declarations

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.

Author 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 sequencing data were submitted in NCBI Database having BioProject ID PRJNA770106 and details of software used are in Table S3. Biological materials used in this study available from the corresponding author.

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 that they have no competing interests.

Author details

[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.

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. 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.*2*022, 36(1):745–752.
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 genome2016, 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. Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok S, Wicker T, Radchuk V, Dockter C, Hedley P, Russell J, et al. A chromosome conformation capture ordered sequence of the barley genome. Nature. 2017;544(7651):427–33.
16. Sætre C, Eroukhmanoff F, Rönkä K, Kluen E, Thorogood R, Torrance J, Tracey A, Chow W, Pelan S, Howe K et al. A Chromosome-Level Genome Assembly of the Reed Warbler (Acrocephalus scirpaceus).Genome Biol Evol2021, 13(9).

17. 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.
18. 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. Sci (New York NY). 2017;356(6333):92–5.
19. Teh B, Lim K, Yong C, Ng C, Rao S, Rajasegaran V, Lim W, Ong C, Chan K, Cheng V, et al. The draft genome of tropical fruit durian (Durio zibethinus). Nat Genet. 2017;49(11):1633–41.
20. Guo C, Wang Y, Yang A, He J, Xiao C, Lv S, Han F, Yuan Y, Yuan Y, Dong X, et al. The Coix Genome Provides Insights into Panicoideae Evolution and Papery Hull Domestication. Mol Plant. 2020;13(2):309–20.
21. Ye C, Wu D, Mao L, Jia L, Qiu J, Lao S, Chen M, Jiang B, Tang W, Peng Q, et al. The Genomes of the Allohexaploid Echinochloa crus-galli and Its Progenitors Provide Insights into Polyploidization-Driven Adaptation. Mol Plant. 2020;13(9):1298–310.
22. 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.
23. 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.
24. Du H, Liang C. Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. Nat Commun. 2019;10(1):5360.
25. Roach M, Schmidt S, Borneman A. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19(1):460.
26. 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.
27. Seppey M, Manni M, Zdobnov E. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in molecular biology (Clifton, NJ)* 2019, 1962:227–245.
28. 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.
29. 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.
30. 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.
31. 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.

32. Vanneste K, Van de Peer Y, Maere S. Inference of genome duplications from age distributions revisited. Mol Biol Evol. 2013;30(1):177–90.
33. 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.
34. 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. Gigascience. 2017;6(11):1–7.
35. 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.
36. Vurture G, Sedlazeck F, Nattestad M, Underwood C, Fang H, Gurtowski J, Schatz M. GenomeScope: fast reference-free genome profiling from short reads. Bioinf (Oxford England). 2017;33(14):2202–4.
37. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinf (Oxford England). 2010;26(5):589–95.
38. 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. Bioinf (Oxford England). 2009;25(16):2078–9.
39. 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.
40. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinf (Oxford England). 2011;27(6):764–70.
41. Koren S, Rhie A, Walenz B, Dilthey A, Bickhart D, Kingan S, Hiendleder S, Williams J, Smith T, Phillippy A. De novo assembly of haplotype-resolved genomes with trio binning.Nat Biotechnol2018.
42. Jiao Y, Peluso P, Shi J, Liang T, Stitzer M, Wang B, Campbell M, Stein J, Wei X, Chin C, et al. Improved maize reference genome with single-molecule technologies. Nature. 2017;546(7659):524–7.
43. Kim D, Langmead B, Salzberg S. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
44. Jarvis D, Ho Y, Lightfoot D, Schmöckel S, Li B, Borm T, Ohyanagi H, Mineta K, Michell C, Saber N, et al. The genome of Chenopodium quinoa. Nature. 2017;542(7641):307–12.
45. 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.
46. 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.
47. Gao L, McCarthy E, Ganko E, McDonald J. Evolutionary history of Oryza sativa LTR retrotransposons: a preliminary survey of the rice genome sequences. BMC Genomics. 2004;5(1):18.
48. Ou S, Jiang N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol. 2018;176(2):1410–22.

49. Haas B, Salzberg S, Zhu W, Pertea M, Allen J, Orvis J, White O, Buell C, Wortman J. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008;9(1):R7.
50. Ou S, Jiang N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob DNA-Uk. 2019;10:48.
51. 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.
52. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li C, Wei L. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 2011;39:W316–322.
53. Chan P, Lin B, Mak A, Lowe T. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 2021;49(16):9077–96.
54. Nawrocki E, Eddy S. Infernal 1.1: 100-fold faster RNA homology searches. Bioinf (Oxford England). 2013;29(22):2933–5.
55. Majoros W, Pertea M, Salzberg S. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinf (Oxford England). 2004;20(16):2878–9.
56. 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.
57. Delcher A, Salzberg S, Phillippy A. Using MUMmer to identify similar regions in large sequence sets.Current protocols in bioinformatics 2003:Unit10.13.
58. 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.
59. Li L, Stoeckert C, Roos D. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89.
60. 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.
61. Nakamura T, Yamada K, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinf (Oxford England). 2018;34(14):2490–2.
62. Höhler D, Pfeiffer W, Ioannidis V, Stockinger H, Stamatakis A. RAxML Grove: an empirical phylogenetic tree database. Bioinf (Oxford England). 2022;38(6):1741–2.
63. Kozlov A, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinf (Oxford England). 2019;35(21):4453–5.
64. Blanc G, Wolfe K. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell. 2004;16(7):1667–78.

65. 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.
66. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 1977;267(5608):275–6.
67. 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.
68. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–91.
69. Lynch M, Conery J. The evolutionary fate and consequences of duplicate genes. Sci (New York NY). 2000;290(5494):1151–5.

Figures

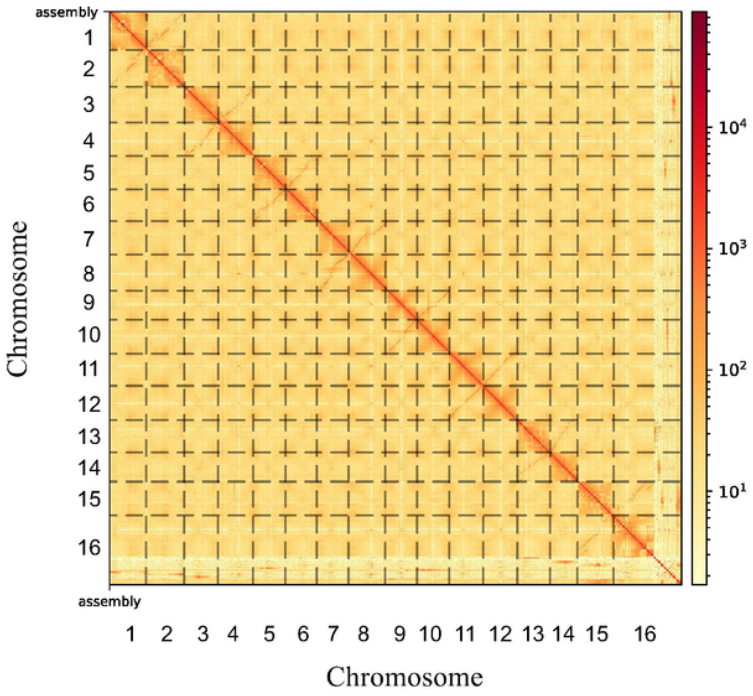

Figure 1

Hi-C-assisted genome assembly of *Trifolium repens*. Hi-C interaction heatmap showing 100-kb resolution super scaffolds

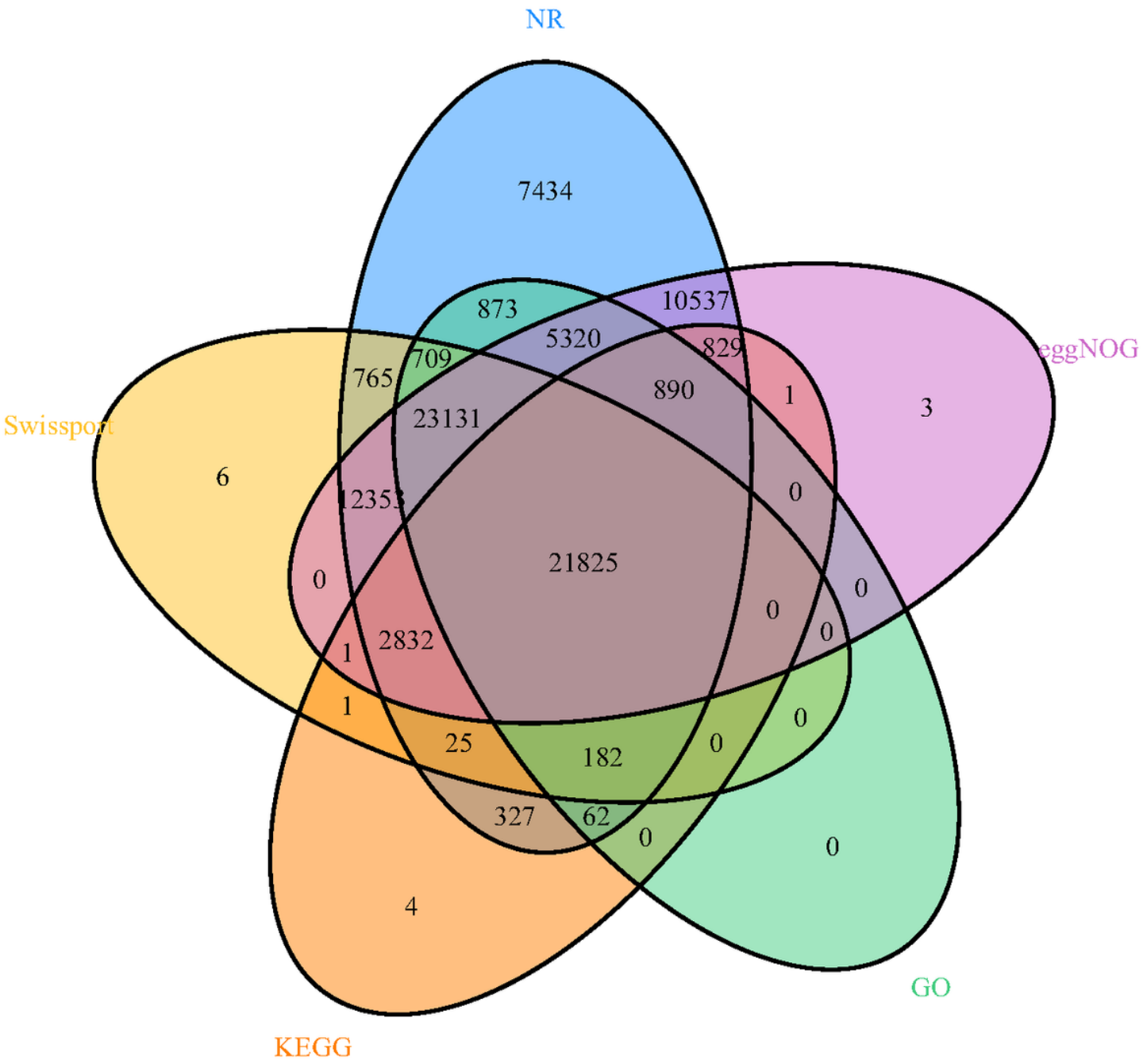

Figure 2

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

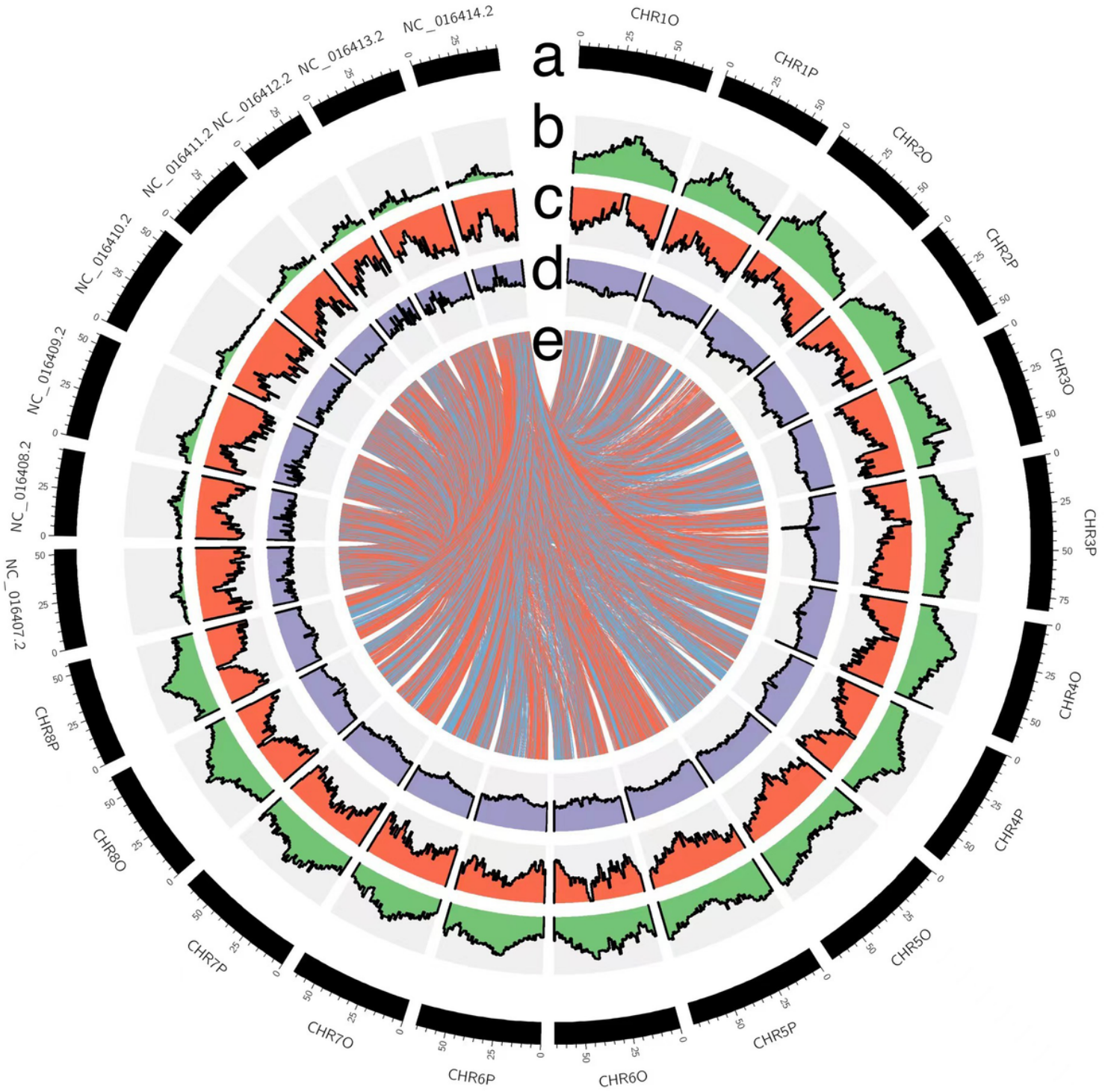

Figure 3

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

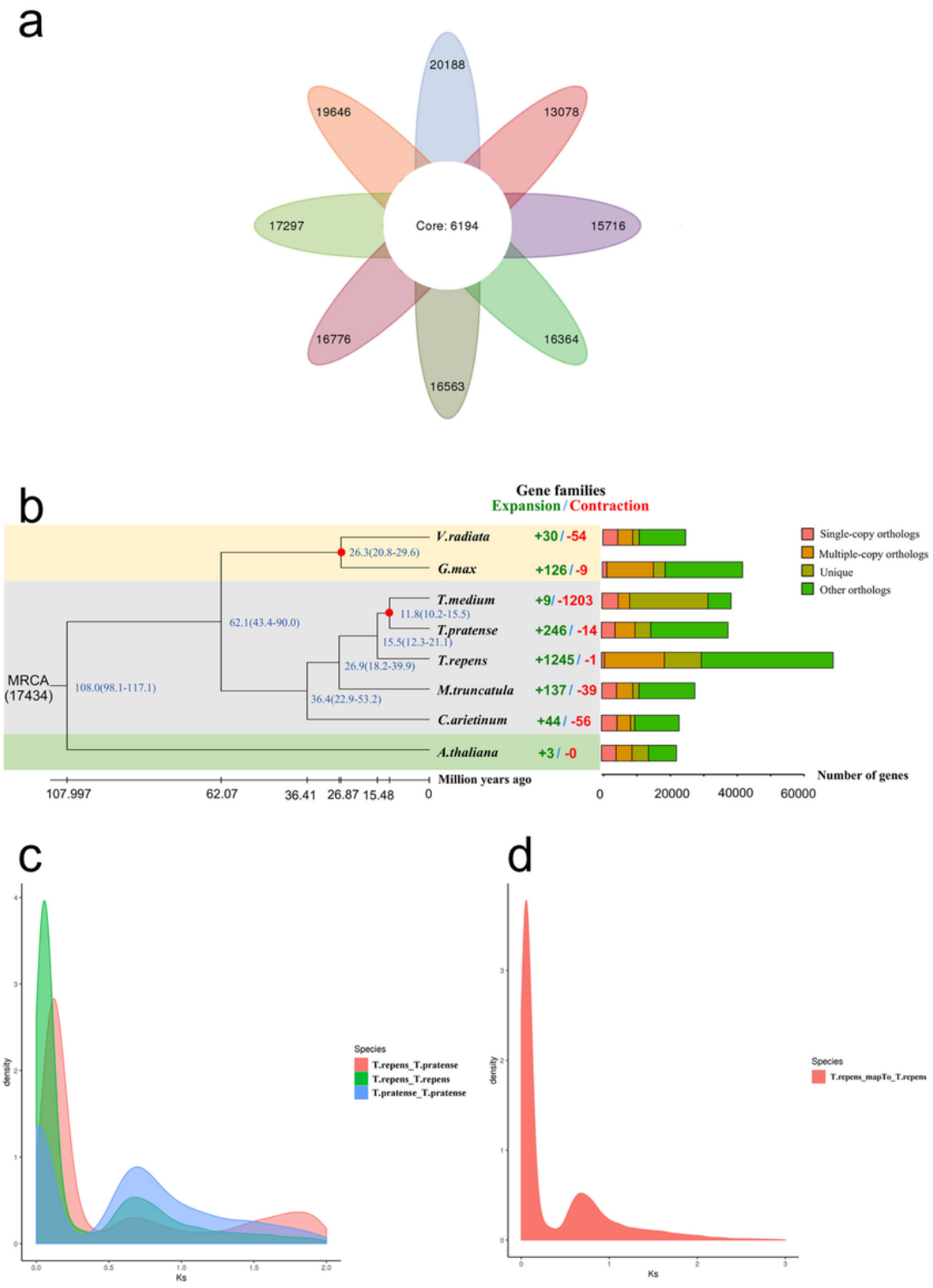

Figure 4

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*

Supplementary Files

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

- TableS1.xls
- TableS2.xls
- TableS3.pdf