BMC Genomic Data

RESEARCH Open Access

Genome assembly of *Melilotus officinalis* provides a new reference genome for functional genomics

Aoran Meng[1], Xinru Li[1], Zhiguang Li[1], Fuhong Miao[1], Lichao Ma[1], Shuo Li[1], Wenfei Sun[1], Jianwei Huang[2] and Guofeng Yang[1*]

Abstract

Background Sweet yellow clover (*Melilotus officinalis*) is a diploid plant (2n = 16) that is native to Europe. It is an excellent legume forage. It can both fix nitrogen and serve as a medicine. A genome assembly of *Melilotus officinalis* that was collected from Best corporation in Beijing is available based on Nanopore sequencing. The genome of *Melilotus officinalis* was sequenced, assembled, and annotated.

Results The latest PacBio third generation HiFi assembly and sequencing strategies were used to produce a *Melilotus officinalis* genome assembly size of 1,066 Mbp, contig N50 = 5 Mbp, scaffold N50 = 130 Mbp, and complete benchmarking universal single-copy orthologs (BUSCOs) = 96.4%. This annotation produced 47,873 high-confidence gene models, which will substantially aid in our research on molecular breeding. A collinear analysis showed that *Melilotus officinalis* and *Medicago truncatula* shared conserved synteny. The expansion and contraction of gene families showed that *Melilotus officinalis* expanded by 565 gene families and shrank by 56 gene families. The contacted gene families were associated with response to stimulus, nucleotide binding, and small molecule binding. Thus, it is related to a family of genes associated with peptidase activity, which could lead to better stress tolerance in plants.

Conclusions In this study, the latest PacBio technology was used to assemble and sequence the genome of the *Melilotus officinalis* and annotate its protein-coding genes. These results will expand the genomic resources available for *Melilotus officinalis* and should assist in subsequent research on sweet yellow clover plants.

Keywords *Melilotus officinalis*, Genome, Assembly, PacBio, HiFi

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

Background

Sweet yellow clover (*Melilotus officinalis*) is a legume that is native to Europe and widely distributed in North America, temperate Europe, the Mediterranean, subtropical Asia, and North Africa [1, 2]. It can adapt to a variety of extreme types of weather, including hot and cold climates. In addition, is tolerant to saline soil. This plant is also used as a forage crop to feed animals and as green manure [3], a source of nectar, and a medicinal herb. This plant also aids in soil and water conservation [4, 5]. Before flowering, the young stems and leaves of the sweet yellow clover plant are easily eaten by animals, and it can be used as silage or converted to grass powder or hay [6]. It is highly nutritious with a crude protein content that is 4.6-fold higher than those of cereals, and the yield and ability of *Melilotus officinalis* to fix nitrogen are also better than those of alfalfa [7]. As a good source of nectar, it secretes large amounts of sugar and is an important source for honeybees to make honey [8]. The honey made from *Melilotus officinalis* is very influential throughout the world and is noted for its clear oral odor, regulation of sleep and metabolism, and enhancement of immunity [9]. It is also used as a medicinal herb and is rich in coumarin, which is an effective treatment for primary lymphedema and the lymphedema associated with radiation therapy, or surgery for breast cancer [10]. *Melilotus officinalis* can also be used to reduce swelling, inflammation, diuresis, and can treat various hemorrhoids and related diseases caused by them [11]. Moreover, it has been used to treat many cancers in recent years [12, 13].

The genomes of two species of *Melilotus* have been reported in recent years, including those of *Melilotus albus* and *Melilotus officinalis* [14, 15]. However, compared with other common legumes, there is limited knowledge on the structural and genetic information of *Melilotus officinalis*, particularly at the genomic level, which has substantially limited its breeding and improvement [16]. In this study, the genome information of *Melilotus officinalis* was obtained by combining Illumina (San Diego, CA, USA), PacBio (Pacific Biosciences of California, Menlo Park, CA, USA), HiFi (High fidelity), and Hi-C (high-throughput chromatin conformation capture) to fully understand the content of its genome and molecular evolutionary history. Hi-C technology was used to observe the collinearity between the chromosomes of *Melilotus officinalis* and its related species. This technique significantly improves the accuracy and sensitivity of evolutionary genetic research and enables the prediction of more robust patterns of genome structure [17]. The purpose of this study was to determine the positive selection of genes and the phylogenetic history of *Melilotus officinalis* when there were historical events and continuous changes in its geographical environment [18]. This study can help with subsequent genomic studies and provide a new research direction to analyze the evolutionary relationship between *Melilotus officinalis* and its close relatives.

Results

Genome survey, sequencing, and assembly

In this study, samples were collected from Best corporation and sequenced using PacBio technology. Several genome parameters of *Melilotus officinalis* were obtained (Fig. 1A). The quality control results revealed that there were 83.6 Gbp of Illumina data with a GC content of approximately 34%. A total of 10,000 read sequences were randomly selected from the filtered clean reads and compared to the NT library through BLASTing, which mapped 97.85% of the sequences. A K-mer analysis can provide a general understanding of the genome before assembly [19]. This K-mer analysis indicated that the genome was 1,080 Mbp, and there were 67.3% repeat sequences and 1.76% heterozygous sequences. PacBio HiFi and Illumina technology were used to sequence the genome of *Melilotus officinalis* [20, 21]. Compared with traditional second-generation sequencing (NGS), the third-generation sequencing (TGS) technology developed by PacBio has the advantages of not requiring PCR amplification, producing long read lengths, and lacking a preference for GC [22, 23]. High-quality HiFi reads after CCS processing with 3 Mbp for HiFi reads, 16 kbp for N50, 57Gbp for base numbers, and a sequencing depth of 52X. A phased string graph was constructed using Hifiasm software, and contigs were generated according to the overlap map. The genome was 1,066 M, and it contained 492 contigs. Contig N50 was 5 Mbp. The largest contig size was 21 Mbp, and the average GC content was 35.38% (Table 1). The Illumina reads were compared with the DNA library to evaluate the quality and completeness of the assembly. The comparison indicated that 85.87% of the properly mapped reads were obtained. The completeness assessment of the assembled genomes was conducted by Benchmarking Universal Single-Copy Orthologs (BUSCOs) and the software TBLASTN, AUGUSTUS, and HMMER. The result was a complete BUSCOs of 96.4%, which showed that the genome assembly was high-quality [24, 25].

Scaffold construction and curation

Hi-C is an extension of chromosome conformation capture (3 C) technology [26, 27]. Hi-C technology has become the primary choice for chromosome-level genome assembly and is widely applied in the assembly of animal and plant genomes [28, 29]. The Hi-C technique was used to obtain 136 Gb of data. A total of 97.34% of the initial assembly based on the PacBio data was scaffolded into eight chromosomes by the Hi-C data. The results of a Hi-C-assisted assembly revealed a

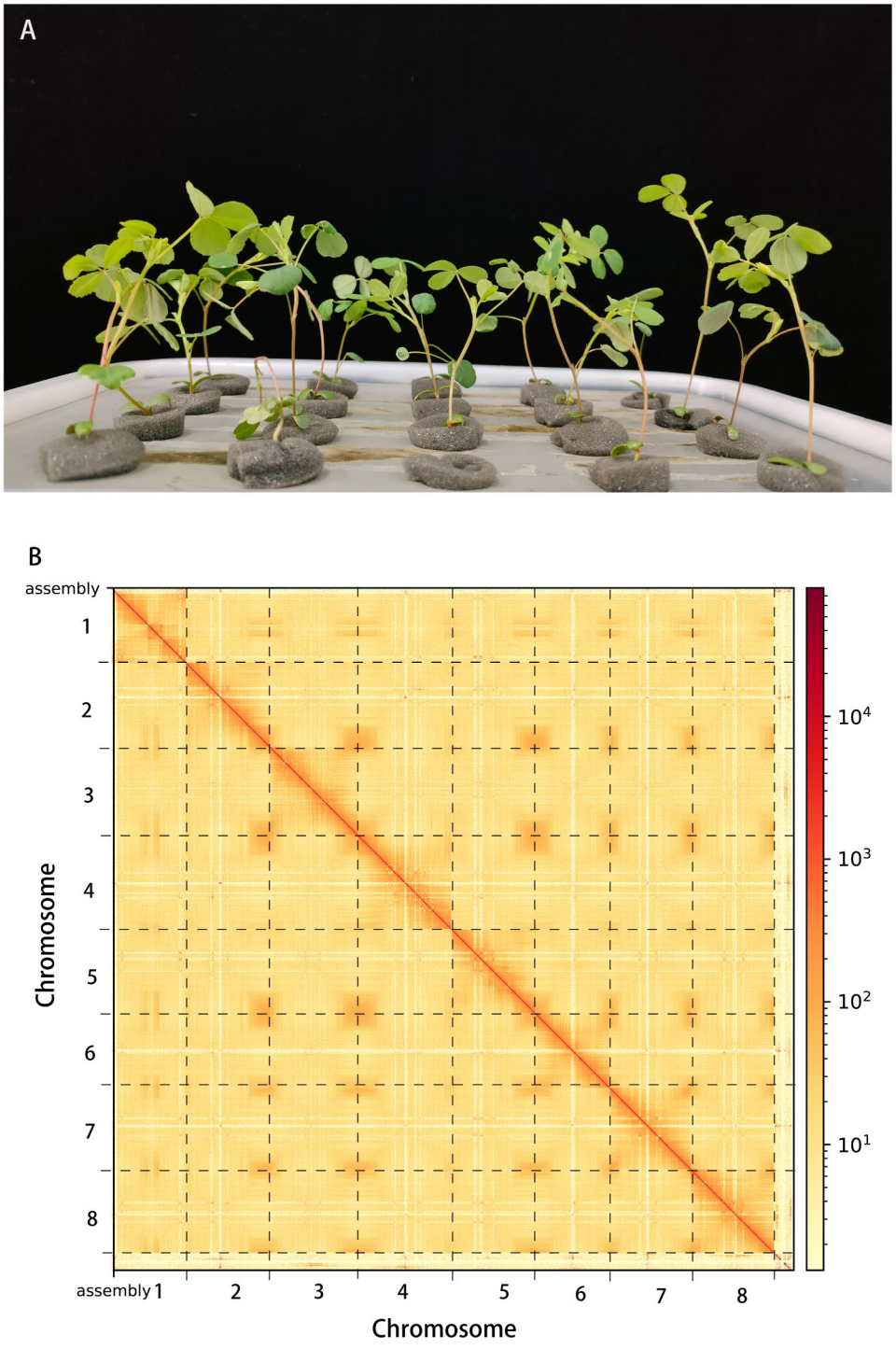

Fig. 1 Plant morphology and Hi-C-assisted genome assembly of sweet yellow clover. (**A**) Phenotype of the sequenced sweet yellow clover plant. (**B**) Hi-C interaction heatmap showing 100-kb resolution super scaffolds. Hi-C, high-throughput chromatin confirmation system

genome size of 1,066 Mbp and a scaffold N50 of 130 Mbp [30]. After the Hi-C-assisted assembly had been completed, the inter-chromosomes and intra-chromosomes exchange interactions required calculation to determine if they were consistent with the principle of Hi-C genome assembly. The linkages within the chromosomes were much stronger than those between the chromosomes. Moreover, the linkages of chromosomes in a close physical location were much stronger than those in a distant physical location (Fig. 1B). These findings suggest that the assembly result was correct. Table 1 summarizes the information on assembly.

Table 1 Summary statistic for the *Melilotus officinalis* genome

	Assembly	This genome
Genome assembly	Estimated genome size	1080 Mbp
	Total length of assembly	1066 Mbp
	Number of contig	492
	Contig N50	5Mbp
	Largest contig	21Mbp
	Number of scaffolds	74
	Scaffold N50	130Mbp
	Chromosome coverage	97.34%
	GC content of genome	35.38%
Annotation		
Transposable elements	Total	761Mbp(71.35%)
	Retrotransposon	569Mbp(53.67%)
	DNA Transposon	62Mbp(5.90%)
Noncoding RNAs	rRNAs	5,168
	tRNAs	1,231
	miRNAs	416
	snRNAs	2,719
Gene models	Number of genes	47,873
	Mean gene length	4,041 bp
	Mean coding sequences length	1,673 bp

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

Organism	Number of genes	Mean coding sequences length(bp)	Exons per transcript	Mean exon length(bp)	Mean intro length(bp)
Melilotus officinalis	47,873	1,673	4.9	367	588
Medicago truncatula	38,823	1,342	4.8	278	540
Trifolium medium	28,496	560	1.7	321	499
Vigna radiata	30,878	1,337	5.1	261	530
Trifolium subterraneum	40,697	1,273	4.4	290	565

Genome annotation

In this study, a total of 71.5% of the genome sequence was identified as repetitive, and it was 49.8% as long as the terminal repeat (LTR) transposable elements [31, 32]. There were 16.09% and 19.96% LTR retrotransposons of Copia and Gypsy, respectively, and there were 15,490 simple repeats in the assembled genome. There were 13 types of noncoding RNA (ncRNA) that totaled 10,016. We obtained 47,873 high confidence gene models by RNA-Seq assembly and gene prediction. The gene models were unevenly distributed on eight chromosomes. The average gene length was 4,041 bp, and each gene contained an average of 4.9 exons. The average lengths of coding sequences, exons, and introns were 1,673 bp, 367 bp, and 588 bp, respectively. We also compared *Melilotus officinalis* with four related species, including *Medicago truncatula* (MtrunA17r5.0-ANR from NCBI), zigzag clover (*Trifolium medium*) (ASM349008v1 from NCBI), mung bean (*Vigna radiata*) (ver6 from NCBI), and subterranean clover (*Trifolium subterraneum*) (TSUd_r1.1 from NCBI). *Melilotus officinalis* had the largest number of genes (47,873) and the longest average coding sequences at 1,673 bp. In contrast, *Trifolium medium* had the fewest genes (28,496). *T. medium* had the fewest average coding sequences and the average number of exons contained in each transcript among these five species. Although there was difference in the number of genes in the remaining three species, the lengths of their mean coding sequences were similar (Table 2). A functional annotation comparison analysis of the five databases annotated 46,776 genes, and the five databases collectively annotated 10,060 genes (Fig. 2). A total of 1,097 genes were not annotated (Table S1).

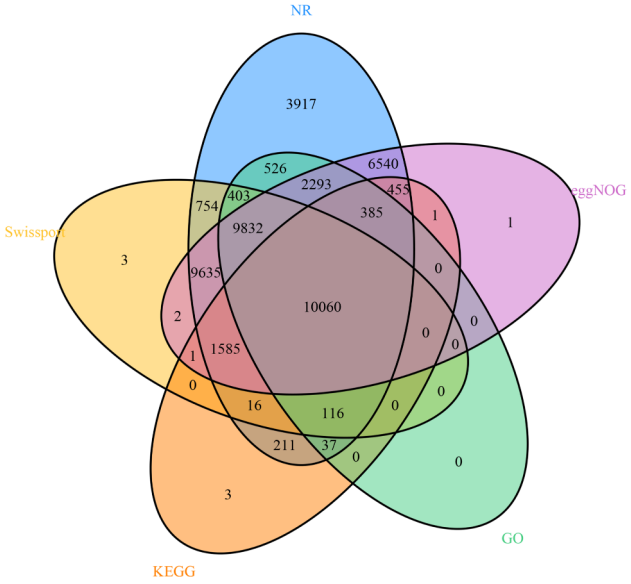

Fig. 2 A Venn diagram that shows the overlap of the five major databases (NR, Swiss-Prot, eggNOG, GO, KEGG) that contain information from the annotation of gene function. GO, Gene Ontology; Kyoto Encyclopedia of Genes and Genomes

Gene family and evolutionary analysis

The relationship between the eight chromosomes possessed by *Medicago truncatula* and *Melilotus officinalis* indicates that the chromosome synteny is conserved between the two species (Fig. 3) [33]. A gene family analysis of the genome of *Melilotus officinalis* and eight common species showed that 39,909 genomes were clustered in 25,207 gene families. There were 28,185 gene families in *T. repens*, and it shared 5,800 of these families among these several species (Fig. 4A). The analysis showed that it had expanded to 565 gene families and contracted 56

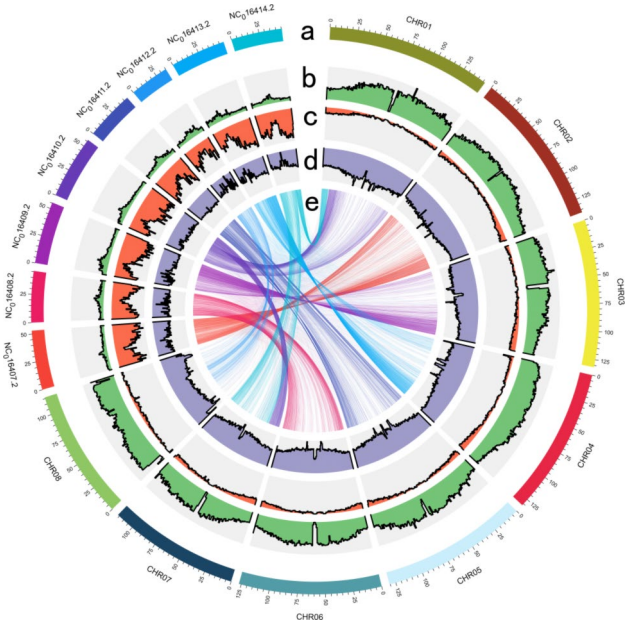

Fig. 3 Feature of the *Medicago truncatula* and *Melilotus officinalis* genome. (**a**) Length of each pseudochromosome (Mbp). (**b**) Distribution of repetitive sequences (%). (**c**) Distribution of gene density (%). (**d**) Distribution of the GC content (%). (**e**) *Medicago_truncatula* and *Melilotus officinalis* synteny analysis; the beginning of NC represents the chromosome of *Medicago truncatula*, while the beginning of CHR represents the chromosome of *Melilotus officinalis*

gene families during the course of evolution. A Gene Ontology (GO) analysis showed that the expanded gene family was related to response to stimulus, nucleotide binding, and small molecule binding. Gene families with biological process are the most abundant (Table S2). These gene families could be involved in plant metabolic processes that are involved in the resistance of plants to stress, which enables the plants to more effectively adapt to changes in the environment. A phylogenetic tree was constructed based on 3,870 single-copy homologous genes, with maize (v. 5.0 from NCBI) as the outgroup. *Melilotus officinalis* clustered with soybean (v. 4.0 from NCBI), chickpea (ASM33114v1 from NCBI), mung bean (v. 6 from NCBI), subterranean clover (TSUd_r1.1 from NCBI), *Medicago truncatula* (MtrunA17r5.0-ANR from NCBI), white clover (AgR_To_v5 from NCBI) and zigzag clover (ASM349008v1 from NCBI) to form a monophyletic group. Single-copy genes of each species were selected as reference markers for the species with incomplete evolutionary studies. The closest relationship was between sweet yellow clover and *Medicago truncatula*, with an estimated time of divergence of approximately 14.4 million years ago (Fig. 4B). Whole-genome duplication (WGD) events are an important indicator of plant evolution and a driving force for the adaptation of plants to various environments [34, 35]. The evolutionary history of the yellow sweet clover plant can be understood by studying the number of synonymous substitutions that occurred at each synonymous site in its genome [36]. The data suggest that both sweet yellow clover and white clover in the self-comparisons had peaks at approximately 0.75 (Fig. 4C) [37, 38]. In addition, the WGD event occurred when the KS value of sweet yellow clover was 0.75 (Fig. 4D).

Discussion

The genomic information of leguminous plants with good agronomic traits is very important for the study of genomics and functional omics [39, 40]. This is not the first report of the assembly of *Melilotus officinalis* since a chromosome-scale assembly of *Melilotus officinalis* has been reported [14, 41]. *Melilotus officinalis* is not only an excellent forage crop that is highly valuable nutritionally; it is also highly valuable medicinally. Although the genome of *Melilotus officinalis* has been published, the differences in sequence were not the same as that observed in this study. This study can enrich the genetic information database of *Melilotus officinalis* and lay a foundation for the further excavation of special genetic markers of *Melilotus officinalis*. For example, if a pan-genome analysis is conducted, it is necessary to sequence as many representative samples of the same plant as possible. Simultaneously, this study adopted the current mainstream Hi-Fi sequencing technology to improve the assembly quality and compensate for gaps in the study of the characteristics of *Melilotus* species. This study enriches the knowledge about legumes and provides research experience for the subsequent study of *Melilotus officinalis.* In this study, a K-mer analysis was used to estimate the genome size, heterozygosity, and repeat sequence ratio, which was the same method utilized in the recently published genome. The K-mer analysis showed that the *Melilotus officinalis* genome was heterozygous (1.76%), highly repetitive (67.3%), and comprised a large and complex genome. The result of K-mer analysis was the same as the published genome with heterozygosity (0.06%) and repetition (71.94%). The genome size was estimated to be 1080 M, which was similar to the genome of *Melilotus officinalis* (1.09 Gb) [14]. Compared with the previously published genome of *Melilotus officinalis*, both the latest PacBio third generation HiFi assembly and sequencing strategies were used to obtain the genome information. The previously published genome assembly of the *Melilotus officinalis* size was 976.27 M (contig N50=7.02 Mbp, scaffold N50=125 Mbp, number of contigs=295) compared with 1,066 M (contig N50=5Mbp, scaffold N50=130 Mbp, number of contigs=492) that was reported this study, which indicated that the quality had significantly improved (Table 1). The assembled genome had an average GC content of 35.38%, which was close to that of the previously assembled *Melilotus*

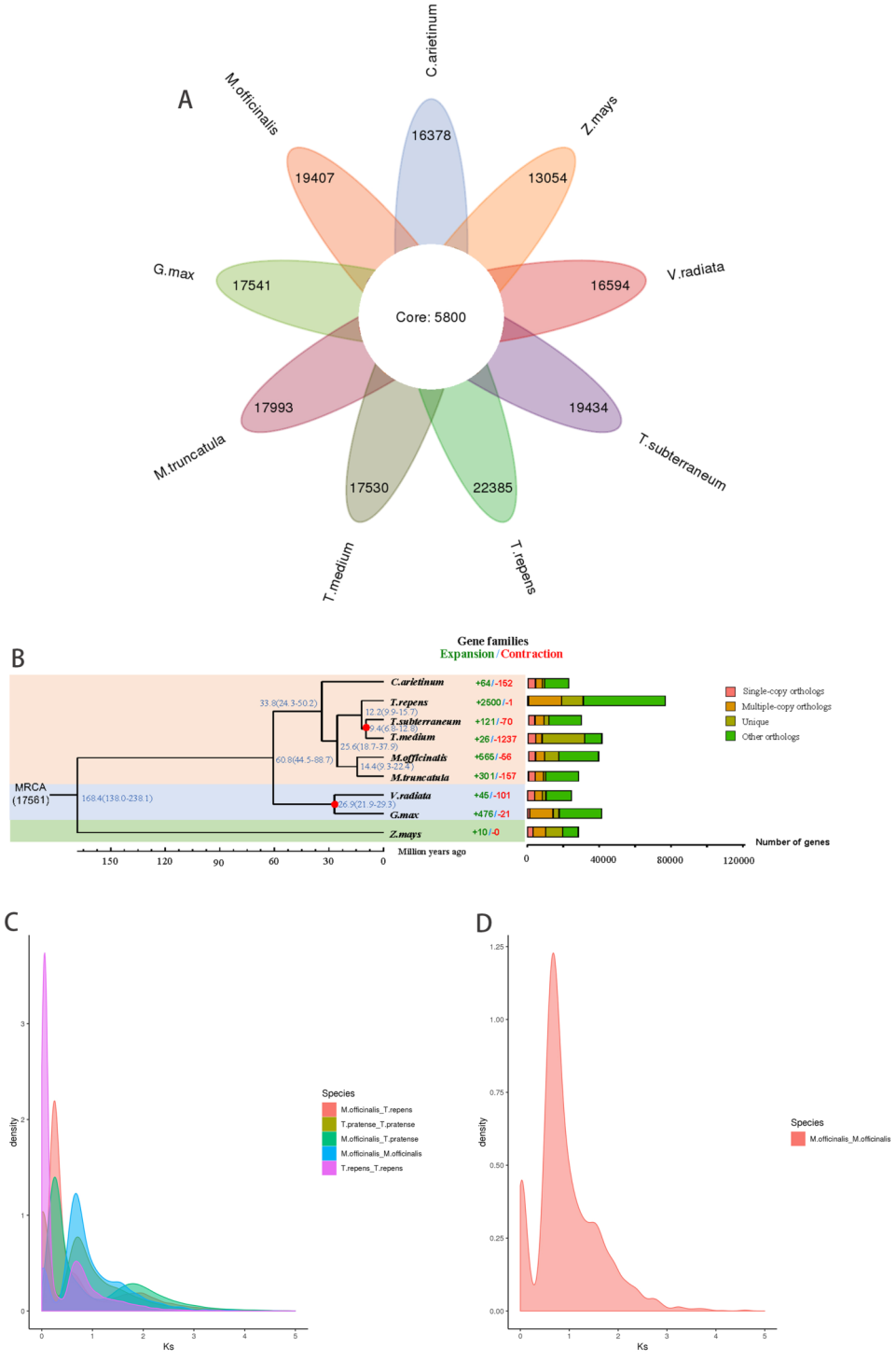

Fig. 4 Gene family clustering and phylogenetic tree analyses of *Melilotus officinalis* and other representative plant genomes. (**A**) A Venn diagram of the number of shared gene families. (**B**) A phylogenetic tree based on shared single-copy gene families (left), gene family expansions and contractions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

officinalis genome (35.50%). The assembled genome at the chromosome level covered 97.34%, which was also close to the previously published genome assembly of *Melilotus officinalis.* BUSCOs were used to compare the published genomes, which enabled an assessment of the integrity of the genomes, and the results showed that both genomes were fully assembled. Compared to the result of the Hi-C-assisted assembly of the published

assembly genome, there were more clean reads and clean bases than in the published assembly genome. The number of unique mapped read pairs was also significantly higher than that of the published assembly genome. After Hi-C scaffolding, the genome annotation revealed 47,873 high-confidence gene models, which was close to the published assembled genome that identified 50,022 annotated genes. Compared with the published assembly genome, the process used to predict the repeat sequences was basically the same, but MITE Hunter v1.0, LTR Finder v1.07 and LTR harvest were used to predict these repeat sequences, which had not been reported in the published assembly genome (Table S3). The results of the prediction of noncoding RNAs revealed that were 5,168 rRNAs, 1,231 tRNAs, 416 miRNAs, and 2,719 snRNAs compared with the published assembly genome (rRNAs=673, tRNAs=934, miRNA=125, snRNAs=244). The prediction of noncoding RNAs in this study resulted in better results than those in the published assembly genome (Table 1). The mean coding sequence length of our assembly genome was 1,673 bp, which was slightly higher than the mean coding sequence length of the published genome (1290.3 bp). However, there were fewer mean exon lengths and mean intro lengths than in the published assembly genome (Table 2). In this study, a Venn analysis was performed on the five major databases to obtain the results of gene function annotation, while a Venn analysis was not performed in the published assembly genome (Fig. 2). The recently published assembly genome used OrthoFinder to analyze clustering of the protein family to compare *Melilotus officinalis* with this parameter in seven other common legumes. These analyses revealed that 891 gene families were unique to *Melilotus officinalis*, and 7,596 gene families were shared by eight legumes. However, in this study, OrthoMCL was used to perform a cluster analysis and compared *Melilotus officinalis* with other eight common legumes. The results showed that *Melilotus officinalis* had 19,407 gene families, and nine legumes shared 5800 gene families. The different results may be owing to comparisons of different species or the use of different software for the analysis. The published assembly genome revealed that there were 635 significantly expanded gene families and 729 significantly contracted gene families in *Melilotus officinalis*. In this study, expansion and contraction of the gene families showed that *Melilotus officinalis* expanded by 565 gene families and shrank by 56 gene families. The expanded genes were mainly involved in proteolysis, peptidase activity and defense response (Figure S1). The contracted genes were mainly involved in response to stress, nuleotide binding, small molecule binding and response to stimulus (Figure S2). These results revealed the expanded and contracted genes affected the stress resistance of *Melilotus officinalis*. These findings provide a basis for subsequent research on molecular breeding. The evolutionary history of *Melilotus officinalis* has been less well-studied. A genome collinearity analysis revealed that *Melilotus officinalis* and *Medicago truncatula* had a high degree of genome collinearity. The WGD events revealed that sweet yellow clover diverged after mung bean, maize (*Zea mays*), soybean (*Glycine mas*), and chickpea (*Cicer ariantum*) and before subterranean clover, white clover, and zigzag clover. Sweet yellow clover and *Medicago truncatula* basically differentiated at the same rate. The ancestors of these species were similar to those of sweet yellow clover. In this study, the genomes of sweet yellow clover and related species were compared at the genomic level. The structural genomic features and gene function of sweet yellow clover were explained by collinear analysis. In addition, a phylogenetic tree construction and analysis, cluster analysis of gene protein families, and a gene contraction and expansion analysis of sweet yellow clover were also conducted. There were also limitations to this study. First, a Hi-C assisted genome assembly was adopted in this study, and the latest T2T genome assembly can complete telomere to telomere assembly, which can improve the quality of genome assembly [42]. Secondly, genomic information is the basis of research function, but the genetic mechanism of related phenotype shape is very complex. How to effectively conduct multi-omics research is also a problem [43]. The future research direction of this area should be to sequence the transcriptome and metabolome of sweet yellow clover, which can facilitate a better understanding of its biological processes [15]. Perhaps single-cell sequencing could be conducted, and single-cell sequencing can help to better understand the regulatory mechanisms of the gene and study the molecular mechanisms at the single-cell level [44]. The genomic information of sweet yellow clover will help to understand the evolution of leguminous plants. The medicinal value of sweet yellow clover also merits study because it produces coumarin [45, 46].

Conclusions

This study reported the third generation Hi-Fi assembly of the PacBio platform, a genome with high coverage and higher completeness in published genomes. Moreover, this study can provide insight into the evolution of vegetation and provide a genetic gene pool for subsequent studies.

Materials and methods

DNA isolation and sequencing

A DNA secure kit (TianGen, Beijing, China) was used to isolate genomic DNA from the leaves of sweet yellow clover in the Grassland Agri-Husbandry Research Center, College of Grassland Science, Qingdao University

(Qingdao, China). The DNA was sequenced by Berry Hekang (Beijing, China) using the PacBio third generation HiFi assembly sequencing platform [20]. First, the quality of samples was tested. The libraries were established to be subjected to PE sequencing using Illumina NovaSeq. Raw reads that contained adapters, duplicates, and low sequence quality were first filtered and then followed by a random selection of 10,000 of the reads for comparison with the NT (Nucleotide Sequence Database) [47] library using BLAST v. 2.12.0 [48]. No significant external contamination was detected. The K-mer counting method was used to estimate the genome size. Clean reads from the Illumina library were used to estimate the genome size using k-mer=23 analysis by Jellyfish v1.1.11 (Table S3) [49]. The formula for estimating the size of the sweet clover genome is as follows:

$$G = K_{number}/K_{depth}$$

where K_{depth} is the expected depth of the k-mers.

Genome assembly and quality evaluation

A NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) was used to determine the quality of the genomic DNA. The degree of DNA degradation and the presence of RNA contamination were analyzed by pulsed field electrophoresis and Fragment Analyzer capillary electrophoresis [50]. The purified genome was subsequently constructed into a SMRTbell library and then sequenced using PacBio SMRT technology [51, 52]. An Agilent 2100 bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) was used to determine the size of library. The data obtained was filtered and then processed using the SMRTLink v. 8.0 software for CCS processing. Hifiasm v. 0.15.2 software was used for assembly, followed by de-hybridization of the contig sequence using the purge-dups v. 1.2.3 software (Table S3) [21, 53]. A single-copy orthologous gene library eudicots_odb10 [54] that was combined using TBLASTN [55], AUGUSTUS v3.2.2 [56], and HMMER [57] software was finally used to evaluate the integrity of the assembled genome [58] (Table S3).

Hi-C data analysis and chromosome construction

Leaf cells (100 mg) of *Melilotus officinalis* were first treated with the cell crosslinking agent paraformaldehyde for 15 min. The DNA and protein were crosslinked; the conformation of DNA was fixed, and glycine was added to prevent the chromatin from crosslinking. The treated leaf tissue was collected and frozen in liquid nitrogen. The leaf tissue was then ground in preparation for subsequent DNA extraction [59]. Biotin was added at the time of end repair to label the oligonucleotide ends. The extracted DNA was subsequently resolved into 350 bp fragments using Covaris and sequenced in the PE150 mode using an Illumina NovaSeq 6000 sequencing platform [60]. The raw reads obtained by sequencing were not all effective, and these raw reads needed to be finely filtered to obtain effective high-quality clean reads. A total of 10,000 read sequences were randomly selected from these filtered clean reads, and their contamination was assessed by alignment to the NT library using BLAST v. 2.12.0 (Table S3) [61]. The Hi-C data were aligned to the preliminary assembled genome using Juicer v. 1.6.2 (Table S3) [28]; the results were filtered and corrected, and then the Hi-C library results were analyzed using 3D DNA 180,922 software (Table S3) [62, 63]. The scaffold of *Melilotus officinalis* was obtained at the chromosome level.

Repeat annotation and gene annotation

The predicted repeats and known repeats in the genome were first masked using Repeat Masker v. 4.1.0 (Table S3) [64, 65]. The repeats were again predicted using MITE Hunter v. 1.0, LTR harvest, LTR Finder v. 1.07, LTR retriever v. 2.8.2, and Repeat Modeler v. 2.0 (Table S3) [66]. The MITEs and LTR transposable elements were then identified using structural prediction methods [67, 68]. Class II transposable element MITEs, as well as nonautonomous transposable elements<2 kb long, were searched from the genome using MITE-Hunter v. 1.0, and the analysis was performed using the software default parameters to enable the prediction of MITEs [69, 70]. The prediction of LTR transposable elements required the use of LTR harvest and LTR Finder v. 1.07 [71]. First, an LTR harvest was used to predict the LTR-RT in the genome using the parameters of the software is-similar 90-vic 10-seed 20-seqids yes-minlenltr 100-maxlenltr 7000-mintsd 4-maxtsd 6-motif TGCA-motifmis 1 [50]. The LTR-RT was predicted using the LTR-Finder v. 1.07 with the following software parameters: -D 15,000-d 1000-L 7000-l 100-p 20-C-M 0.9 [72] (Table S3). The repeats in masked genomes were identified *de novo* using RepeatModeler v. 2.0 with the following software parameters: -engine ncbi-pa 60 [73]. RepeatMasker v. 4.1.0 was then used to block the repetitive sequences in the genome, and the software utilized the following parameters: -s-nolow-norna-gff-engine ncbi-parallel 20 [74] (Table S3). The ab initio prediction for tRNA was performed using the software tRNAscan-SE v. 2.0, and rRNA and other types of ncRNA were searched by their similarity when aligned with the Rfam database (https://ftp.ebi.ac.uk/pub/databases/Rfam/14.1/) [75, 76] (Table S3). All the repetitive regions except tandem repeats were soft-masked to annotate the proteins that encoded genes [77, 78]. GeMoMa-1.6.1 was used to compare the protein sequences of related species with the assembled genomes. These comparisons were then combined with the comparison of RNA data and assembly results

to obtain exon and intron boundary information and improve the prediction accuracy [79, 80]. A comprehensive transcriptome database was constructed using PASA (v. 2.0.1). The gene structure was predicted with AUGUSTUS v3.2.2 combined with the RNA-Seq data, SNAP v6.0 and GlimmerHMM v3.0.4 [81, 82] (Table S3). The RNA-seq data was used to annotate the gene structure to optimize the accuracy of gene structure annotation and provide a reliable training set for the *de novo* prediction software. The parameters were trained with the training set, and the Scaffold with the masked repeat sequence was utilized [83, 84] The predictions obtained using these packages were combined using EVIdenceModeler (EVM) r2012-06-25 [68] (Table S3), and then 36,511 genes were retrieved and functionally annotated by BLAST searches against databases, including NR (http://ftp.ncbi.nlm.nih.gov/blast/db/) [85], Swiss-Prot (http://ftp.ebi.ac.uk/pub/databases/uniprot/knowledgebase/uniprot_sprot.fasta.gz) [86], eggNOG (http://eggnog6.embl.de.) [87], Gene Ontology (GO) (http://geneontology.org/) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/) [88]. A Venn diagram of the five major databases was then performed to obtain more accurate information on the functional annotation of the genes [89].

Comparative analysis

Genomic collinearity was analyzed on *Melilotus officinalis* and its related species *Medicago_truncatula* using MuMMER v. 4.1 software, The parameters of the software were as follows: nucmer-g 1000-c 90-l 200 [90] (Table S3). Nine protein families were identified by OrthoMCL cluster analysis, including chickpea, soybean, *M. truncatula*, zigzag clover, white clover [*T. repens*], subterranean clover, mung bean, maize, and *Melilotus officinalis*) [91]. An all-vs-all BLAST alignment of all the sequences of *Melilotus officinalis* genes that encode proteins (with $1e^{-5}$ as the default e-value) was first performed and then followed by a calculation of the sequence similarity [92, 93]. The Markov clustering algorithm was then used for cluster analysis with an expansion coefficient of 1.5 to obtain the clustering results for the protein families [94, 95]. Owing to a lack of research on the evolution of *Melilotus officinalis*, selected species single-copy genes were used as a reference marker to select the four degenerate sites to construct a supergene using the MAFFT v. 7.310 software for multiple sequence alignment [96, 97] (Table S3). The most suitable base substitution model was selected with RAxML software that was based on the maximum likelihood (ML) species phylogenetic tree. MCMCtree was used from the PAML v. 4.9e package based on a single copy gene family (parameter: burn-in =5,000,000, sample-number=1,000,000, and sample-frequency=50) [98] (Table S3). The time of differentiation was estimated. Time calibration points (correction points) were from the Timetree website [99]. The gene family was then analyzed using CAFE v. 3.1 software and GO functional enrichment analysis for the genes in these families (Table S3) [100–102]. A branch-site model can detect positive selection that occurs in a particular clade and only affects a portion of the locus. The one-to-one orthologous proteins were selected from *Melilotus officinalis* and its related species, and the homologous protein sequences were aligned using the default parameters of PRANK software [103, 104].

gBlocks were used to filter the alignment results with the parameters -t=c-e=. For ft-b4=5-d=y, the CODEML test in PAML v. 4.9e was located in a specific clade and only affected positive selection at certain sites. It was corrected for multiple hypothesis testing using the Chi2 program in PAML v. 4.9e. The main parameter was 2 degrees of freedom [105] (Table S3).

The WGD events were detected using the duplicate age distribution method. The longest protein sequences of genes in the *Melilotus officinalis* genome were then aligned using BLASTP. The alignment was filtered using the DAG chainer, and the synonymous substitution rate was calculated using the Yn 00 tool in the PAML v. 4.9e software package [106]. A map of density distribution based on the Ks values of all the paralog gene pairs and the Ks values of orthologous gene pairs between the genomes of *Melilotus officinalis*, white clover, and other related species was then drawn using MATLAB [107] (Table S3).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-024-01224-y.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Acknowledgements

The authors would like to thank Professors Guofeng Yang, Zengyu Wang, and Juan Sun (Professor of Grassland Science, Qingdao Agricultural University) for their help in analyzing the data and writing the manuscript. We are also grateful for the research funding provided by the College of Grassland Science of Qingdao Agricultural University and the experimental help provided by Berry Hekang (Beijing, China). We would like to thank MogoEdit (https://www.mogoedit.com) for its English editing during the preparation of this manuscript.

Author contributions

AM and GY conceived and designed this research. AM analyzed the data and wrote the manuscript. AM, JH, XL, ZL and LM analyzed the data. LS participated in the discussion of the results. LM, WS, FM and ZL collected samples. GY contributed to the evaluation and discussion of the results and

revisions of the manuscript. All the 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), the First Class Grassland Science Discipline Program of Shandong Province (1619002), China and the Foundation Project of Shandong Natural Science Foundation (ZR2022MC031).

Data availability
All raw data were submitted in NCBI Database (SRR23985850, SRR23985849, SRR23985851, SRR23985848). The details of software used are in Table S3. and the genome assembly and annotation were uploaded in the dedicated public repositories (assembly of *Melilotus officinalis*: DOI: 10.6084/m9.figshare.23590107, genome annotation of Melilotus officinalis: DOI: 10.6084/m9.figshare.23590161).

Declarations

Ethics approval and consent to participate
Sweet yellow clover is not an endangered or protected species in China, and it was purchased from BEST grass industry and planted in a light incubator. The seeds were collected by Professor Guofeng Yang in the BEST grass industry. All the study procedures were conducted in accordance with the relevant guidelines.

Consent for publication
Not applicable.

Competing interests
Jianwei Huang was employed by Berry Genomics Corporation, Beijing, China. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 2 January 2024 / Accepted: 10 April 2024
Published online: 18 April 2024

References

1. Baimiev A, Gubaidullin II, Baimiev A, Cheremis AV. Effects of natural and hybrid lectins on the legume-rhizobium interactions. Prikl Biokhim Mikrobiol. 2009;45(1):84–91.
2. Huang R, Snedden WA, diCenzo GC. Reference nodule transcriptomes for Melilotus officinalis and Medicago sativa cv. Algonquin. Plant Direct. 2022;6(6):e408.
3. Mouad LB, Ahmed A, Abdelaziz J, Mohamed A, Mohammed A. Effect of yellow sweetclover (Melilotus officinalis) hay compared with Lucerne (Medicago sativa) hay on carcass characteristics and meat quality of male goat kids. J Adv Vet Anim Res. 2022;9(4):617–24.
4. Chen J, Bird GW, Mather RL. Impact of multi-year cropping regimes on Solanum tuberosum Tuber yields in the Presence of Pratylenchus penetrans and verticillium dahliae. J Nematol. 1995;27(4S):654–60.
5. Robson DB, Knight JD, Farrell RE, Germida JJ. Ability of cold-tolerant plants to grow in hydrocarbon-contaminated soil. Int J Phytorem. 2003;5(2):105–23.
6. Atwood SS. Cytogenetics and breeding of forage crops; sweet clover. Adv Genet. 1947;1:55–7.
7. Puntillo M, Gaggiotti M, Oteiza JM, Binetti A, Massera A, Vinderola G. Potential of lactic acid Bacteria isolated from different forages as silage inoculants for improving Fermentation Quality and Aerobic Stability. Front Microbiol. 2020;11:586716.
8. Jasicka-Misiak I, Makowicz E, Stanek N. Polish yellow Sweet Clover (Melilotus officinalis L.) Honey, Chromatographic fingerprints, and Chemical markers. Molecules 2017, 22(1).
9. Frunze O, Brandorf A, Kang EJ, Choi YS. Beekeeping Genetic Resources and Retrieval of Honey Bee Apis mellifera L. Stock in the Russian Federation: a review. Insects 2021, 12(8).
10. Ilhan M, Ali Z, Khan IA, Kupeli Akkol E. A new isoflavane-4-ol derivative from (Melilotus officinalis). Nat Prod Res. 2019;33(13):1856–61.
11. Paun G, Neagu E, Albu C, Savin S, Radu GL. In Vitro evaluation of antidiabetic and anti-inflammatory activities of polyphenolic-rich extracts from Anchusa officinalis and Melilotus officinalis. ACS Omega. 2020;5(22):13014–22.
12. Parvizpour S, Masoudi-Sobhanzadeh Y, Pourseif MM, Barzegari A, Razmara J, Omidi Y. Pharmacoinformatics-based phytochemical screening for anticancer impacts of yellow sweet clover, Melilotus officinalis (Linn.) Pall. Comput Biol Med. 2021;138:104921.
13. Pitaro M, Croce N, Gallo V, Arienzo A, Salvatore G, Antonini G. Coumarin-Induced Hepatotoxicity: a narrative review. Molecules 2022, 27(24).
14. He Q, Li Z, Liu Y, Yang H, Liu L, Ren Y, Zheng J, Xu R, Wang S, Zhan Q. Chromosome-scale assembly and analysis of Melilotus officinalis genome for SSR development and nodulation genes analysis. Plant Genome 2023:e20345.
15. Wu F, Duan Z, Xu P, Yan Q, Meng M, Cao M, Jones CS, Zong X, Zhou P, Wang Y, et al. Genome and systems biology of Melilotus albus provides insights into coumarins biosynthesis. Plant Biotechnol J. 2022;20(3):592–609.
16. Zhou L, Hou F, Wang L, Zhang L, Wang Y, Yin Y, Pei J, Peng C, Qin X, Gao J. The genome of Magnolia Hypoleuca provides a new insight into cold tolerance and the evolutionary position of magnoliids. Front Plant Sci. 2023;14:1108701.
17. Bouwman BAM, Crosetto N, Bienko M. The era of 3D and spatial genomics. Trends Genet. 2022;38(10):1062–75.
18. Rice ES, Green RE. New approaches for Genome Assembly and Scaffolding. Annu Rev Anim Biosci. 2019;7:17–40.
19. Armstrong J, Fiddes IT, Diekhans M, Paten B. Whole-genome alignment and comparative annotation. Annu Rev Anim Biosci. 2019;7:41–64.
20. Benevenuto J, Ferrao LFV, Amadeu RR, Munoz P. How can a high-quality genome assembly help plant breeders? *Gigascience* 2019, 8(6).
21. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170–5.
22. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356(6333):92–5.
23. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin CS, et al. Improved maize reference genome with single-molecule technologies. Nature. 2017;546(7659):524–7.
24. Jiao WB, Schneeberger K. The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol. 2017;36:64–70.
25. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol Biol. 2019;1962:227–45.
26. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE. 2014;9(11):e112963.
27. Zhang H, Wang Y, Deng C, Zhao S, Zhang P, Feng J, Huang W, Kang S, Qian Q, Xiong G, et al. High-quality genome assembly of Huazhan and Tianfeng, the parents of an elite rice hybrid tian-you-Hua-Zhan. Sci China Life Sci. 2022;65(2):398–411.
28. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing Loop-Resolution Hi-C experiments. Cell Syst. 2016;3(1):95–8.
29. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36(9):2896–8.
30. Jauhal AA, Newcomb RD. Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. Mol Ecol Resour. 2021;21(5):1416–21.
31. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.
32. Zavallo D, Crescente JM, Gantuz M, Leone M, Vanzetti LS, Masuelli RW, Asurmendi S. Genomic re-assessment of the transposable element landscape of the potato genome. Plant Cell Rep. 2020;39(9):1161–74.
33. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40(7):e49.
34. Kumar S, Nei M, Dudley J, Tamura K. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform. 2008;9(4):299–306.
35. Vanneste K, Van de Peer Y, Maere S. Inference of genome duplications from age distributions revisited. Mol Biol Evol. 2013;30(1):177–90.
36. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000;290(5494):1151–5.

37. Yan Z, Sang L, Ma Y, He Y, Sun J, Ma L, Li S, Miao F, Zhang Z, Huang J, et al. A de novo assembled high-quality chromosome-scale Trifolium pratense genome and fine-scale phylogenetic analysis. BMC Plant Biol. 2022;22(1):332.
38. Wang H, Wu Y, He Y, Li G, Ma L, Li S, Huang J, Yang G. High-quality chromosome-level de novo assembly of the Trifolium repens. BMC Genomics. 2023;24(1):326.
39. Hamilton JP, Buell CR. Advances in plant genome sequencing. Plant J. 2012;70(1):177–90.
40. Michael TP, VanBuren R. Building near-complete plant genomes. Curr Opin Plant Biol. 2020;54:26–33.
41. Wibberg D, Blom J, Ruckert C, Winkler A, Albersmeier A, Puhler A, Schluter A, Scharf BE. Draft genome sequence of Sinorhizobium meliloti RU11/001, a model organism for flagellum structure, motility and chemotaxis. J Biotechnol. 2013;168(4):731–3.
42. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44–53.
43. Jamil IN, Remali J, Azizan KA, Nor Muhammad NA, Arita M, Goh HH, Aizat WM. Systematic Multi-omics Integration (MOI) Approach in Plant systems Biology. Front Plant Sci. 2020;11:944.
44. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50(8):1–14.
45. Bazazzadegan N, Dehghan Shasaltaneh M, Saliminejad K, Kamali K, Banan M, Khorram Khorshid HR. The effects of Melilotus officinalis Extract on expression of Daxx, Nfkb and Vegf genes in the Streptozotocin-Induced Rat Model of sporadic Alzheimer's Disease. Avicenna J Med Biotechnol. 2017;9(3):133–7.
46. Zhang J, Di H, Luo K, Jahufer Z, Wu F, Duan Z, Stewart A, Yan Z, Wang Y. Coumarin Content, Morphological Variation, and Molecular Phylogenetics of Melilotus. Molecules 2018, 23(4).
47. Arita M, Karsch-Mizrachi I, Cochrane G. The international nucleotide sequence database collaboration. Nucleic Acids Res. 2021;49(D1):D121–4.
48. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
49. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–70.
50. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
51. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol 2018.
52. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37(10):1155–62.
53. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
54. Waterhouse RM, Zdobnov EM, Kriventseva EV. Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. Genome Biol Evol. 2011;3:75–86.
55. Gertz EM, Yu YK, Agarwala R, Schaffer AA, Altschul SF. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol. 2006;4:41.
56. Hoff KJ, Stanke M. Predicting genes in single genomes with AUGUSTUS. Curr Protoc Bioinf. 2019;65(1):e57.
57. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res. 2018;46(W1):W200–4.
58. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. Genome Project Data Processing S: the sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
59. Ramirez F, Bhardwaj V, Arrigoni L, Lam KC, Gruning BA, 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.
60. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 2020;30(9):1291–305.
61. Quinlan AR. BEDTools: the swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinf. 2014;47(12):11.
62. van Dijk M, Bonvin AM. 3D-DART: a DNA structure modelling server. Nucleic Acids Res 2009, 37(Web Server issue):W235–239.
63. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19(1):460.
64. Hou XG, Zhang X, Guo DL. Identification and analysis methods of plant LTR retrotransposon sequences. Yi Chuan. 2012;34(11):1491–500.
65. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80.
66. Storz G. An expanding universe of noncoding RNAs. Science. 2002;296(5571):1260–3.
67. Edgar RC, Myers EW. PILER: identification and classification of genomic repeats. Bioinformatics. 2005;21(Suppl 1):i152–158.
68. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 2008;9(1):R7.
69. Tarailo-Graovac M. N Chen 2009 Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinf Chap. 4 41011–141014.
70. Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. 2010;38(22):e199.
71. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 2008;18(12):1979–90.
72. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59.
73. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):955–64.
74. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 2007, 35(Web Server issue):W265–268.
75. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucleic Acids Res. 2003;31(1):439–41.
76. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004;20(16):2878–9.
77. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 2004, 32(Web Server issue):W20–25.
78. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–5.
79. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 2011, 39(Web Server issue):W316–322.
80. 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.
81. Ouyang S, Buell CR. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res. 2004;32(Database issue):D360–363.
82. Nachtweide S, Stanke M. Multi-genome Annotation with AUGUSTUS. Methods Mol Biol. 2019;1962:139–60.
83. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 2004;32(Web Server issue):W309–312.
84. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011;27(17):2325–9.
85. Reau M, Lagarde N, Zagury JF, Montes M. Nuclear receptors database including negative data (NR-DBIND): a database dedicated to nuclear receptors binding data including negative data and pharmacological Profile. J Med Chem. 2019;62(6):2894–904.
86. Gasteiger E, Jung E, Bairoch A. SWISS-PROT: connecting biomolecular knowledge via a protein database. Curr Issues Mol Biol. 2001;3(3):47–55.
87. Hernandez-Plaza A, Szklarczyk D, Botas J, Cantalapiedra CP, Giner-Lamia J, Mende DR, Kirsch R, Rattei T, Letunic I, Jensen LJ, et al. eggNOG 6.0: enabling comparative genomics across 12 535 organisms. Nucleic Acids Res. 2023;51(D1):D389–94.
88. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
89. Zdobnov EM, Apweiler R. InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001;17(9):847–8.
90. Bailey JA, Eichler EE. Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet. 2006;7(7):552–64.
91. Li L, Stoeckert CJ Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89.

92. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. 2001;11(6):1005–17.
93. Blanc G, Wolfe KH. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell. 2004;16(7):1667–78.
94. Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 2002;19(6):908–17.
95. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noel 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.
96. Delcher AL, Salzberg SL, Phillippy AM. Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinformatics* 2003, Chap. 10:Unit 10 13.
97. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
98. Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res. 2005;15(8):1153–60.
99. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 2005;22(12):2472–9.
100. Hahn MW, Han MV, Han SG. Gene family evolution across 12 Drosophila genomes. PLoS Genet. 2007;3(11):e197.
101. Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. Nat Rev Genet. 2020;21(7):428–44.
102. Abramova A, Osinska A, Kunche H, Burman E, Bengtsson-Palme J. CAFE: a software suite for analysis of paired-sample transposon insertion sequencing data. Bioinformatics. 2021;37(1):121–2.
103. Loytynoja A. Phylogeny-aware alignment with PRANK. Methods Mol Biol. 2014;1079:155–70.
104. Jammali S, Djossou A, Ouedraogo WDD, Nevers Y, Chegrane I, Ouangraoua A. From pairwise to multiple spliced alignment. Bioinform Adv. 2022;2(1):vbab044.
105. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al. Ancestral polyploidy in seed plants and angiosperms. Nature. 2011;473(7345):97–100.
106. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 1977;267(5608):275–6.
107. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–45.

Publisher's Note