frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 07 March 2023
DOI 10.3389/fpls.2023.1124536

Check for updates

OPEN ACCESS

EDITED BY
Yiqiang Zhao,
China Agricultural University, China

REVIEWED BY
Quanzi Li,
Chinese Academy of Forestry, China
Sha Chen,
Key Laboratory of Beijing for Identification and Safety Evaluation of Chinese Medicine, Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China

*CORRESPONDENCE
Jingyuan Song
✉ jysong@implad.ac.cn

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 15 December 2022
ACCEPTED 02 February 2023
PUBLISHED 07 March 2023

CITATION
Gao L, Xu W, Xin T and Song J (2023) Application of third-generation sequencing to herbal genomics. *Front. Plant Sci.* 14:1124536. doi: 10.3389/fpls.2023.1124536

Application of third-generation sequencing to herbal genomics

Longlong Gao, Wenjie Xu, Tianyi Xin and Jingyuan Song*

Key Lab of Chinese Medicine Resources Conservation, State Administration of Traditional Chinese Medicine of the People's Republic of China, Engineering Research Center of Chinese Medicine Resource of Ministry of Education, Institute of Medicinal Plant Development, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China

There is a long history of traditional medicine use. However, little genetic information is available for the plants used in traditional medicine, which limits the exploitation of these natural resources. Third-generation sequencing (TGS) techniques have made it possible to gather invaluable genetic information and develop herbal genomics. In this review, we introduce two main TGS techniques, PacBio SMRT technology and Oxford Nanopore technology, and compare the two techniques against Illumina, the predominant next-generation sequencing technique. In addition, we summarize the nuclear and organelle genome assemblies of commonly used medicinal plants, choose several examples from genomics, transcriptomics, and molecular identification studies to dissect the specific processes and summarize the advantages and disadvantages of the two TGS techniques when applied to medicinal organisms. Finally, we describe how we expect that TGS techniques will be widely utilized to assemble telomere-to-telomere (T2T) genomes and in epigenomics research involving medicinal plants.

KEYWORDS

third-generation sequencing, PacBio, nanopore, herbal genomics, medicinal plant, molecular identification

1 Introduction

There is a long history of traditional medicine use. In 2015, Chinese scientist Youyou Tu won the Nobel Prize for her outstanding contribution to the discovery of artemisinin, refreshing the global perception of traditional Chinese medicine. There is little doubt that many further health-promoting discoveries will be made by studying traditional medicine. Over recent decades, many phytochemical and pharmacological research projects have investigated the bioactive components and underlying mechanisms of herbal medicine. However, the available genetic information on herbal medicines long remained lacking due to the high cost of the predominant first-generation sequencing technique, Sanger sequencing. It was not until the emergence of next-generation sequencing (NGS) that the situation improved. Because of its high-throughput and low costs, NGS has made it affordable for most researchers to sequence the genomes and transcriptomes of medicinal plants, greatly promoting the development of herbal genomics. Nevertheless, as research developed, the inherent shortcomings of NGS—especially the short read lengths—became a new bottleneck hindering the development of herbal genomics.

When using NGS techniques, the characteristics of short read lengths make it difficult to assemble the raw fragments into high-quality contigs or scaffolds, especially those with high heterozygosity or a high proportion of repeat sequences. Yet, the development of third-generation sequencing (TGS) has brought us a great opportunity to solve these problems. However, TGS is still unfamiliar to many researchers. Therefore, here we introduce the principles, pipelines, and sequencing instruments of two mainstream TGS techniques, PacBio single-molecule real-time (SMRT) sequencing technology and Oxford Nanopore technology (ONT). We compare these two techniques with Illumina, the predominant NGS technique. To demonstrate how TGS can be applied to herbal genomics, we have chosen several classic studies of genomics, transcriptomics, and molecular identification as examples to dissect the specific processes and summarize the advantages and disadvantages of TGS when applied in medicinal organisms. This work will provide a meaningful reference for traditional medicine and genomic researchers.

2 Insights into main TGS

The first single-molecule sequencing technology was developed by Helicos Bioscience, but it is rarely used now because it is comparatively time-consuming and has short read lengths (~32 bp) (Harris et al., 2008; Orlando et al., 2011). Currently, there are two widely used TGS technologies, PacBio SMRT technology and ONT.

2.1 PacBio SMRT technology

2.1.1 The principle of SMRT

Similar to Illumina sequencing, SMRT is based on the principle of sequencing-by-synthesis, acquiring sequence information during the amplification process of nucleic acid molecules. Before sequencing, both ends of the targeted double-stranded DNA (dsDNA) molecule are ligated with hairpin adapters to form dumbbell-shaped templates (i.e., SMRTbells). These adapters allow DNA polymerases and primers to bind with the SMRTbells (Figure 1A). After binding, the SMRTbells are sequenced on a SMRT cell chip. There are thousands of Zero-model waveguides (ZMWs) lined up on each SMRT cell. The ZMWs limit the observation volume to avoid the influence of the fluorescence of uncombined deoxyribonucleoside triphosphates (dNTPs), which allows the detection of a single dNTP. Once the SMRTbells are loaded onto the SMRT cell, a proportion of them fall into the ZMWs. Then, the SMRTbells are fixed on the bottom of ZMWs through the interaction between the biotin on the polymerase and the streptavidin on the glass plate of the ZMWs (Eid et al., 2009). The DNA polymerases catalyze the continuous incorporation of dNTPs labeled by different fluorophores into complementary strands (Flusberg et al., 2010). When the polymerases capture the labeled dNTPs, they emit distinct fluorescence pulses under excitation light (Figure 1B). Four classic dNTP types can be recognized from their featured fluorescence signature (Eid et al., 2009).

The methylated bases in the DNA template change the incorporation kinetics of polymerases, which enables SMRT to directly detect methylated bases without chemical modifications (Flusberg et al., 2010). Moreover, because of the circular structure of the SMRTbell and the replacement sequencing ability of the DNA polymerase, inserted DNA templates can be repeatedly sequenced,

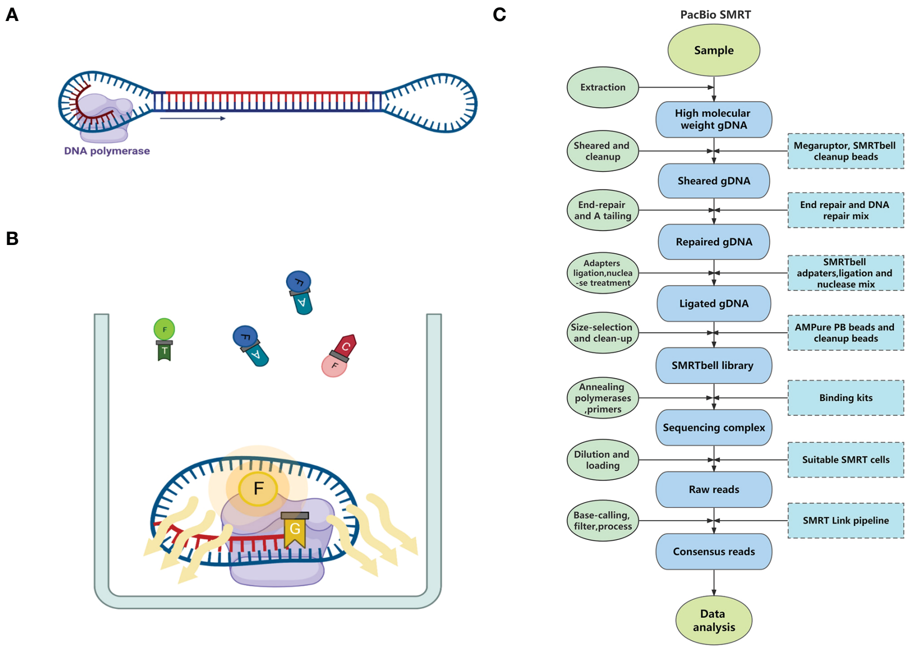
A
DNA polymerase
B
F
T
A
C
G
C
PacBio SMRT
Sample
Extraction
High molecular weight gDNA
Sheared and cleanup
Megaruptor, SMRTbell cleanup beads
Sheared gDNA
End-repair and A tailing
End repair and DNA repair mix
Repaired gDNA
Adapters ligation,nuclea -se treatment
SMRTbell adpaters,ligation and nuclease mix
Ligated gDNA
Size-selection and clean-up
AMPure PB beads and cleanup beads
SMRTbell library
Annealing polymerases ,primers
Binding kits
Sequencing complex
Dilution and loading
Suitable SMRT cells
Raw reads
Base-calling, filter,process
SMRT Link pipeline
Consensus reads
Data analysis

FIGURE 1
The principle and pipeline of PacBio SMRT Technology. **(A)** A dumbbell-shaped template for sequencing (SMRTbell) consists of adapters in both ends, a double-stranded DNA template, and a DNA polymerase. **(B)** The process of sequencing in the Zero-model waveguide (ZMW). **(C)** The pipeline of PacBio SMRT sequencing.

yielding many copies of both template and complementary strands. Aligning these copies greatly improves the sequencing accuracy. This sequencing strategy is also known as circular consensus sequencing (CCS) (Travers et al., 2010).

2.1.2 The pipeline of SMRT

Sample and library preparation: In the genome sequencing of medicinal plants, fresh leaves are often used as samples for DNA extraction. High molecular weight genomic DNA (HMW gDNA) should be extracted from the samples because the amount and length distribution of extracted gDNA are important for subsequent library construction. Usually, high fidelity (HiFi) library construction for whole genome sequencing of plants requires at least 1 μg of DNA input per 1 Gb of genome length. DNA molecules ≥10 kb should account for more than 90%, and molecules ≥30 kb should account for more than 50%. Additionally, when this amount of DNA cannot be extracted from the samples, alternative workflows are available (low DNA input workflow and ultra-low DNA input workflow). After quality control of HMW gDNA molecules, they need to be sheared to suitable sizes by the Megaruptor system, followed by cleanup with SMRTbell cleanup beads. Then, the sheared gDNA undergoes end-repair, A tailing, adapters ligation, and nuclease treatment in a thermocycler, followed by size-selection using AMPure PB Beads or cleanup with SMRTbell cleanup beads to form the SMRTbell library (usually 15–18 kb). Finally, DNA polymerases and primers are annealed to the SMRTbell library using Binding kits (e.g., Sequel II binding kits 3.2), and then the final sequencing complexes are constructed[1].

Sequencing: After dilution, the SMRTbell library is loaded onto PacBio sequencers with one or more SMRT cells, each of which can yield HiFi reads up to 4 Gb in one run. The runtime is flexible depending on the amount of data needed by the assembly[2].

Primary data analysis: Base-calling and primary filtering analysis are performed on the sequencer. SMRT Link, PacBio DevNet, and other software tools are available to process the raw SMRT data[3]. The complete pipeline of SMRT is shown in Figure 1C.

2.1.3 Sequencing instruments of SMRT

There are six long-read sequencing instruments based on SMRT sequencing technology: PacBio RS, RSII, sequel, sequel II, sequel IIe, and the newly released Revio. Among them, PacBio RS is the first sequencer commercialized by PacBio. As an early-released instrument, PacBio RS had a relatively low throughput, short average read lengths (~1.5 kb) and a high error rate (~13%) (Quail et al., 2012). With technological advances, the throughput of PacBio sequencers has increased by several hundred folds while the average read lengths underwent a 10-fold increase. The accuracy has also been improved to more than 99% due to the extensive use of the CCS strategy (Table 1). These instruments have been used in many genomic and transcriptomic studies of medicinal plants, among which PacBio sequel is used frequently, probably because of its high throughput and relatively low costs.

2.2 Oxford nanopore technology

ONT is another popular TGS technique. Unlike previous sequencing technologies, ONT does not detect fluorescence, light, or pH signals. Instead, it distinguishes bases by detecting electrical signals (Clarke et al., 2009).

[10] 1[0] https://www.pacb.com/technology/hifi-sequencing/sequel-system,

[11] 1[1] https://www.pacb.com/revio/

[12] 1[2] https://nanoporetech.com/products

TABLE 1 Comparison of SMRT, ONT and Illumina representative instruments.

Technique	Instrument	Principle	Accuracy	Read length	Throughput per run/Gb	Run time/h	Reference
PacBio SMRT	RS	Sequencing-by-synthesis	~87%	~1.5 kb	0.1	2	(Quail et al., 2012)
	Sequel II		≥99%	~15-18 kb	30	30	PacBio website[10, 11,]
	Revio				360	24	
Oxford Nanopore	MinION	Threading DNA or RNA through nanopore protein	~85%	Equal to the length of input DNA or RNA	50	72	Nanopore website [12, 13,] (Jain et al., 2017)
	GridION				250	72	
	PromethION				14,000	72	
Illumina	Miniseq	Sequencing-by-synthesis	~99.6%	2×150 bp	7.5	4-24	Illumina website[14] (Quail et al., 2012)
	Miseq			2×300 bp	15	4-55	
	Nova-Seq X			2×150 bp	16,000	13-48	

[10] 1[0] https://www.pacb.com/technology/hifi-sequencing/sequel-system,
[11] 1[1] https://www.pacb.com/revio/
[12] 1[2] https://nanoporetech.com/products
[13] 1[3] https://nanoporetech.com/products/kits
[14] 1[4] https://www.illumina.com.cn/systems/sequencing-platforms.html

2.2.1 The principle of ONT

ONT originated from a brand-new idea of threading a single-stranded nucleic acid molecule through a nanopore protein (Clarke et al., 2009) (Figure 2A). The sequencing process of ONT takes place in a container filled with an electrolyte solution. A lipid double-layer membrane embedded with a nanopore protein is placed in the container. Under an applied voltage, a stable current is formed in the nanopore due to the flow of ions. Therefore, when a nucleic acid molecule passes through, the nanopore is partially blocked, and the stable current is interfered with. Because the structures of nucleotides are different, they cause distinct interferences with the current (Figure 2B). For this reason, nucleotides of sequenced nucleic acid molecules can be distinguished from their distinctive current variations; in this way, sequence information can be decoded (Clarke et al., 2009; Jain et al., 2016).

2.2.2 The pipeline of ONT

Sample and library preparation: Library preparation kits for whole genome sequencing, targeted DNA sequencing, and RNA sequencing are all provided by ONT[4]. Here we take genome sequencing on MinION using Ligation Sequencing Kit V14 as an example. The process of library preparation using this sequencing kit takes about 60 min. Fresh leaves are frequently used for DNA extraction in whole genome sequencing by ONT. First, HMW gDNA can be extracted from plant tissues using the NEB Monarch HMW DNA Extraction Kit or other compatible extraction kits. Then, researchers can choose whether to conduct fragmentation or size selection in the pipeline. If not, the yielded read length will equal the input fragment length. Second, the extracted HMW gDNA undergoes DNA repair and end-preparation (end-prep) using NEBNext FFPE DNA Repair Mix and NEBNext Ultra II End Repair/dA-tailing Module reagents. Third, sequencing adapters are ligated to the repaired ends of DNA molecules using a ligation sequencing kit and some other reagents, followed by cleanup, after which the sequencing library is prepared. Finally, about 1 µl of the DNA library is quantified using a Qubit fluorometer, and then the DNA library is produced as 12 µl at 10–20 fmol[5]. In addition, an automated device, VolTRAX, released by ONT enables hands-free and standard sequencing library construction[6].

Sequencing: Different library preparation kits are compatible with different versions of flow cells, which should be confirmed before sequencing. After priming the flow cell, 10–20 fmol of the final DNA library is advised to be loaded onto the flow cell. The time for one run is up to 72 h.

Primary data analysis: Data acquisition is usually performed by MinKNOW, and base-calling can be conducted by MinKNOW, GUPPY, and many other algorithms available on GitHub[7]. The complete pipeline of ONT is shown in Figure 2C.

2.2.3 Sequencing instruments of ONT

Flongle, MinION, GridION, and PromethION are the main sequencing instruments of ONT. Among them, GridION and PromethION are bench-top instruments with high throughput and, therefore, usually used for large-scale sequencing projects[8] (Jain et al., 2016) such as whole genome sequencing of humans, mammal animals, and plants. MinION is a portable instrument weighing only 90 g that can be used for small sequencing projects, such as microorganism genomes and rapid sequencing outside the laboratory (and even in space) (Jain et al., 2016; McIntyre et al., 2016). Flongle is a single-use product generating 1–2 Gb of data, which is suitable for even smaller projects, such as plasmid and viral sequencings[9].

2.3 Comparison between TGS and NGS techniques

There are three main strengths of SMRT and ONT compared to Illumina, the predominant NGS technique (Table 1). First, the average read length of SMRT and ONT (usually ≥10 kb) is much longer than that of Illumina (~150–300 bp). Second, SMRT and ONT have much lower guanine and cytosine-content bias (GC bias) than Illumina and other NGS techniques (Benjamini and Speed, 2012; Roeh et al., 2017; Sato et al., 2019; Castaño et al., 2020). Third, SMRT and ONT enable researchers to directly detect base modifications without any of the special processes needed by Illumina (Flusberg et al., 2010; Simpson et al., 2017; LaBarre et al., 2019). However, it is worth noting that although the accuracy of SMRT can be greatly improved by the CCS strategy (Travers et al., 2010), the error rate of ONT (~15%) (Jain et al., 2017) is still much higher than that of Illumina (~0.4%) (Quail et al., 2012).

2.4 Comparison between ONT and SMRT

Both ONT and SMRT are single-molecule sequencing techniques with long read lengths, low GC bias, and the ability to directly detect base modifications. However, there are many differences between these two techniques. First, the principles of ONT and SMRT are very different. SMRT inherited and developed the basic principle and labeling method of Illumina sequencing, namely sequencing-by-

[4] https://nanoporetech.com/products/kits

[5] https://community.nanoporetech.com/protocols/genomic-dna-by-ligation-sqk-lsk114

[6] https://nanoporetech.com/products/voltrax

[7] https://community.nanoporetech.com/docs/sequence/sequencing_software/data-analysis

[8] https://nanoporetech.com/products

[9] https://nanoporetech.com/products/flongle

[10] 1[0] https://www.pacb.com/technology/hifi-sequencing/sequel-system,

[11] 1[1] https://www.pacb.com/revio/

[12] 1[2] https://nanoporetech.com/products

[13] 1[3] https://nanoporetech.com/products/kits

[14] 1[4] https://www.illumina.com.cn/systems/sequencing-platforms.html

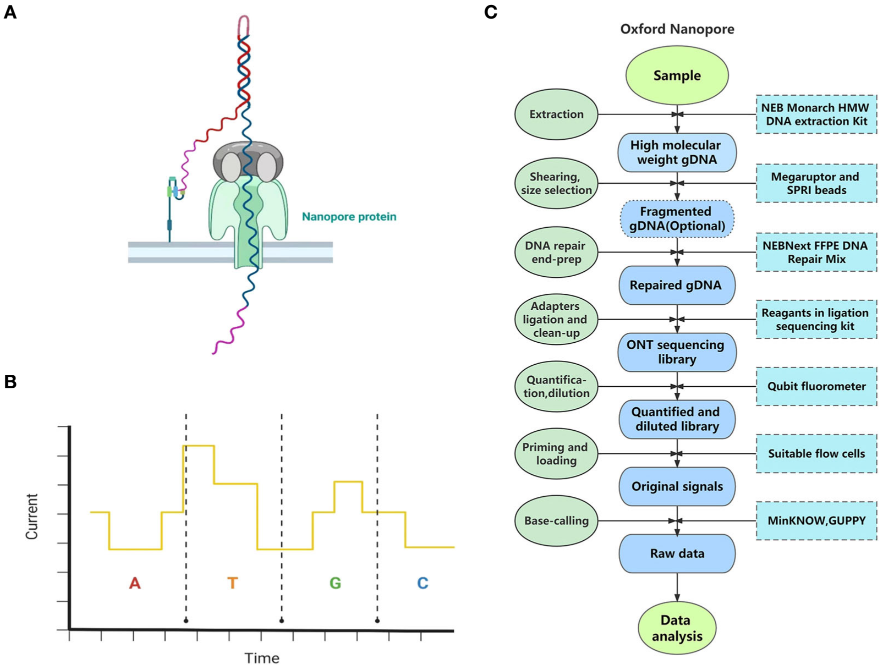

FIGURE 2
The principle and pipeline of Oxford Nanopore Technology. **(A)** The specific scene of threading a single-stranded DNA through a nanopore protein. **(B)** Different dNTPs can cause distinct interference to the current when passing through the nanopore protein (the current variations demonstrated do not represent the true influence of dNTPs.) **(C)** The pipeline of Oxford Nanopore Technology.

synthesis and fluorescence labeling, while ONT is a novel approach that threads the nucleic acid molecules through nanopore proteins and distinguishes nucleotides by electrical signals. Second, the types of instruments of SMRT and ONT are different. Sequencers of SMRT are all bench-top instruments with relatively high throughput, while ONT devices can be either bench-top, high-throughput devices (GridION, PromethION) or portable, comparatively low-throughput devices (MinION, Flongle) are available (Table 1).

3 Application of TGS to herbal genomics

3.1 Decoding whole genomes of medicinal plants

3.1.1 Nuclear genomes

Decoding nuclear genomes of medicinal plants with high heterozygosity and a high proportion of repeat sequences using NGS techniques often leads to fragmented genome assemblies. With much longer read lengths, TGS enables researchers to uncover the genomic regions missed by NGS techniques. To date, more than 100 nuclear genomes of medicinal plants have been sequenced using TGS, most of which are assembled to a chromosome level combined with high-throughput chromosome conformation capture (Hi-C) mapping technology (Cheng et al., 2021c) (Supplementary Table 1). Among them, a recent study of *Gardenia jasminoides* (Figure 3A) demonstrated the classic processes of applying TGS to nuclear genomes research of medicinal plants, from library construction to data analysis. Therefore, it is taken as an example here (Xu et al., 2020b).

In the reference study, ONT, Illumina, and Hi-C are combined to gain a chromosome-level assembly of the *G. jasminoides* genome, while RNA sequencing (RNA-seq) is used to evaluate assembly quality, predict protein-coding genes, and calculate the expression level of genes.

Genome size estimation: Before ONT sequencing, flow cytometry (Pfosser et al., 1995) and *k*-mer distribution analysis (Manekar and Sathe, 2018) were used to estimate genome size and heterozygosity. Based on the two methods, the genome of *Gar. jasminoides* was predicted to have a total size of 550.6 ± 9 Mb and a high heterozygosity of 2.2%, implicating that it is challenging to assemble this genome.

Library preparation: Libraries for ONT, Illumina, RNA-seq, and Hi-C were constructed. In the reference study, the fresh leaves of *Gar. jasminoides* were pooled for DNA extraction of ONT and Illumina sequencing. Seven organs, including fruits at different maturity stages, of *Gar. jasminoides* were collected for RNA-seq and the measurement of crocin content. The Hi-C library was constructed with fresh tissue from *Gar. jasminoides*. For ONT library construction, HMW gDNA was extracted from the pooled leaves and then fragmented, size selected, and purified to get large fragments, after which the large fragments underwent end-prep, adapter ligation, tether attachment, and then an ONT library was constructed.

Sequencing and assembly: The complete ONT library of *Gar. jasminoides* was sequenced on GridION X5, and the raw data was base-called using Guppy (v1.8.5), generating 2.67 Gb reads with an N50 of 21.6 kb. For assembly, the authors developed a satisfactory package (Canu-SMARTdenovo-3×Pilon) by testing various *de novo* assembly pipelines. Specifically, using this package, base-called ONT reads were corrected and trimmed by Canu and then assembled with

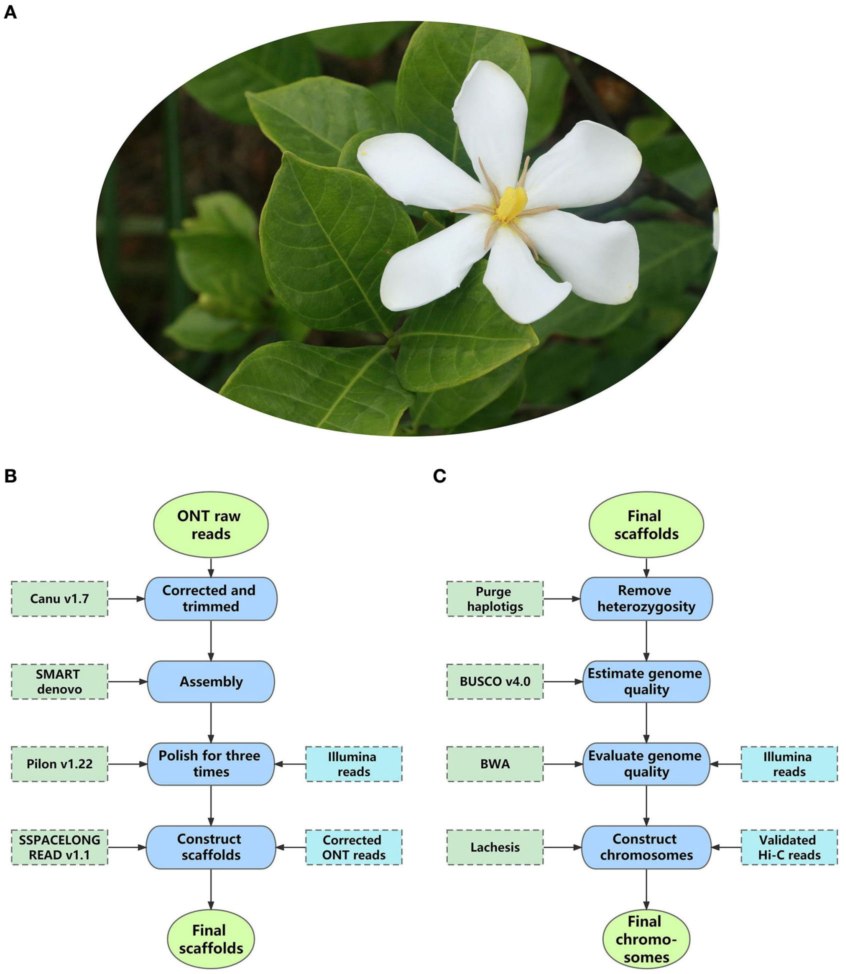

FIGURE 3
The pipeline of chromosome-level assembly of *Gardenia jasminoides* genome. **(A)** The flower of *Gardenia jasminoides*. **(B)** The pipeline for assembling the raw data into scaffolds. **(C)** The pipeline for assembling scaffolds into pseudo-chromosomes.

SMARTdenovo, followed by Illumina short reads to polish the Canu-SMARTdenovo contigs with Pilon three times. The final scaffolds were assembled with the polished contigs, and ONT reads were corrected using Canu; heterozygous sequences were eliminated by Purge Haplotigs. Finally, the researchers acquired a 534.1 Mb assembly with a contig N50 of 1.0 Mb (Figure 3B). The quality of the assembly was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) and mapped with Illumina short reads from the DNA and RNA libraries of *Gar. jasminoides*, respectively, which found 95.0% complete BUSCOs. Hi-C technology was used to further improve the quality of the assembly. As a result, 99.5% of sequences from the assembly were scaffolded into 11 pseudo-chromosomes using the Lachesis package. At this stage, the final chromosome-level genome of *Gar. jasminoides* was successfully constructed. It was 535 Mb in size, with a scaffold N50 of 44 Mb (Figure 3C).

Further data analysis: The first step of data analysis for medicinal plant genomes is usually genome annotation, followed by comparative genomic analysis and expression analysis. In the reference study, at first, the chromosome-level genome of *Gar. jasminoides* was used for genome annotation, including the annotation of repeat elements, the prediction and functional annotation of protein-coding genes, and the annotation of non-coding RNA. This genome was then used for comparative genomic analyses, including synteny analysis between *Gar. jasminoides* and *Coffea canephora*, phylogenetic analysis between *Gar. jasminoides* and ten additional angiosperms, followed by mapping transcriptome reads obtained from RNA-seq of seven organs of *Gar. jasminoides* to the annotated genes to calculate the relative expression level of genes (fragments per kilobase of exon per million reads mapped, FPKM). Furthermore, genome-wide analysis was conducted after mapping, in which genes from three families related to crocin biosynthesis were identified, including 14 carotenoid cleavage dioxygenases (CCDs) genes, 18 aldehyde dehydrogenases (ALDHs)-like genes, and 237 UDP-glucosyltransferases (UGTs) genes. The *Gar. jasminoides* crocin biosynthetic was elucidated after expressing 14 candidate crocin biosynthetic genes in *Escherichia coli* to test their enzymatic activity. In addition, the above-mentioned comparative analysis revealed the evolution of crocin and caffeine biosynthesis genes in Rubiaceae (Figure 4).

In addition to the software used in the reference study, many new tools suitable for TGS techniques have recently been developed. Ratatosk (Holley et al., 2021) was developed for hybrid error correction while CONSENT (Morisse et al., 2021) was designed for self-correction of TGS. For genome assembly, new assemblers are

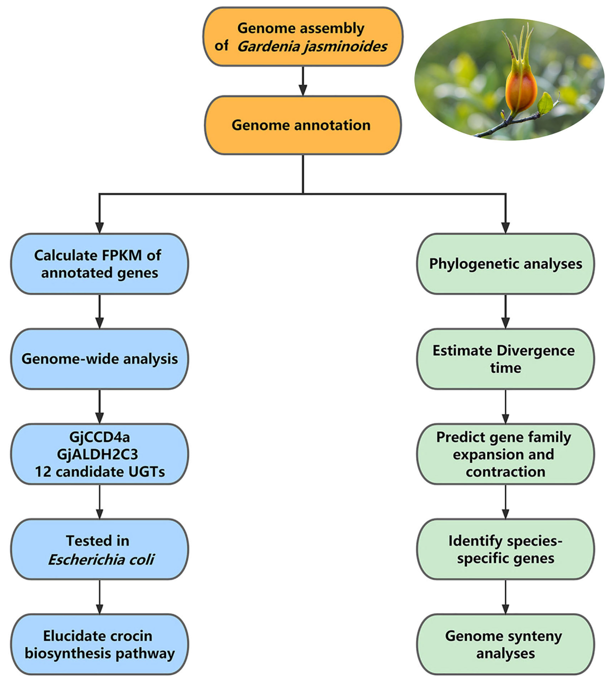
Genome assembly of *Gardenia jasminoides*
Genome annotation
Calculate FPKM of annotated genes
Genome-wide analysis
GjCCD4a
GjALDH2C3
12 candidate UGTs
Tested in *Escherichia coli*
Elucidate crocin biosynthesis pathway
Phylogenetic analyses
Estimate Divergence time
Predict gene family expansion and contraction
Identify species-specific genes
Genome synteny analyses

FIGURE 4
The workflow of data analysis of *Gardenia jasminoides* genome. The data analysis of the *Gardenia jasminoides* genome mainly includes genome annotation followed by comparative analysis and expression analysis of annotated genes.

available, such as hifiasm (Cheng et al., 2021a) and Nextdenovo (https://github.com/Nextomics/NextDenovo). And for scaffolding, RegScaf (Li and Li, 2022) was developed to resolve large genomes and repeat regions, while YaHS (Zhou et al., 2023) is suitable for chromosome-scale scaffold construction using Hi-C data. Regarding annotation, PhyloCSF++ (Pockrandt et al., 2022) is a newly updated tool for differentiating protein-coding and non-coding regions, and TransposonUltimate (Riehl et al., 2022) is a newly developed tool for transposon classification, annotation, and detection. Advanced bioinformatics software have also greatly facilitated the genome research of medicinal plants.

High-quality reference genome assemblies have provided valuable genetic information for investigating the biosynthesis of secondary metabolites, such as triptolide, morphinan, and icaritin. By integrating the genome, transcriptome, and metabolome of *Tripterygium wilfordii*, a cytochrome P450 (CYP728B70) in *T. wilfordii* was identified to catalyze the oxidation of a methyl to the acid moiety of dehydroabietic acid in triptolide biosynthesis, providing clues for elucidating the biosynthetic pathway of triptolide (Tu et al., 2020). *(S)- to (R)-reticuline* (*STORR*) gene fusion is key for morphinan biosynthesis in *Papaver somniferum*. CYP450 and oxidoreductase genes that combined to form the gene fusion were also identified by paralog analysis using a chromosome-level *Papaver somniferum* genome assembly (Guo et al., 2018). In addition, an important flavonoid prenyltransferase (*Ep*PT8) in *Epimedium pubescens* was proven to be involved in the biosynthesis of icaritin and its derivatives by whole genome search using a chromosome-level genome assembly of *E. pubescens* (Shen et al., 2022a). Notably, the biosynthetic pathways of iridoids in *Rehmannia glutinosa* (Ma et al., 2021) and crocin in *Gardenia jasminoides* (Xu et al., 2020b) were successfully elucidated.

The availability of high-quality reference genomes also facilitates the research of the evolutionary history of important gene clusters, biosynthetic pathways, and species in different families. For example, the convergent evolution of CYP82D and CYP706X members in Lamiaceae and Asteraceae (Gao et al., 2022) and the divergent evolution of caffeine and crocin biosynthetic pathways were revealed based on TGS genome assemblies. Information on the evolution of species in Euphorbiaceae (Wang et al., 2021b), Asteraceae (Shen et al., 2018), and Magnoliid (Shang et al., 2020) has been provided as well in recent studies. These are valuable resources for subsequent functional genomics, molecule-assisted breeding, and synthetic biology research (Xin et al., 2019).

Although many chromosome-level assemblies of medicinal plant genomes have been completed, gaps and highly repetitive regions remain to be resolved, such as the centromere and telomere regions. Recently, by combining ONT ultra-long reads, PacBio HiFi reads, and Hi-C technology, researchers successfully obtained truly gapless telomere to telomere (T2T) reference genome assemblies of *Hordeum vulgare* (Navrátilová et al., 2022), *Arabidopsis thaliana* (Wang et al., 2022a), *Citrullus lanatus* (Deng et al., 2022), and several other plants. Compared to chromosome-level assemblies, T2T assemblies are more complete, can be used to discover almost all genomic variations, and enable research into centromere and telomere regions. However, T2T assembly of medicinal plants is still rarely reported.

3.1.2 Organelle genomes

Organelle genomes are also important genetic resources for medicinal plant utilization. To date, the chloroplast genomes (cp-genomes) of more than 20 important medicinal plants (Table 2) have been sequenced using TGS. Further, the complete mitochondrial genomes of several important medicinal plants have been sequenced by TGS (Table 3). Most of the obtained cp-genomes were determined as circular molecules with quadripartite structures consisting of a pair of inverted repeat regions (IRs), a large single-copy region (LSC), and a small single-copy region (SSC), while a few of which were also identified as tripartite or bipartite structures. Here we take the research of *Salvia miltiorrhiza* (Chen et al., 2014) as an example to introduce the achievements of this research and the classic workflow when applying TGS to the sequencing of medicinal plant cp-genomes.

Sample and library preparation: The sample and library preparation of cp-genomes is similar to that of nuclear genomes. In the selected study, fresh leaves were prepared from *S. miltiorrhiza* for gDNA isolation. The gDNA was extracted using a plant genomic DNA kit (Tiangen, China). Libraries consisting of inserted fragments 1 kb and 10 kb in size were prepared and used for subsequent SMRT sequencing.

Sequencing and assembly: The SMRT sequencing of gDNA was conducted under the guidance of the manufacturer provided by PacBio and the raw sequences were preprocessed using the SMRT Analysis workflow. Regarding assembly, first, more than 200 cp-genomes were downloaded and blasted against the cp-genome of *S. miltiorrhiza*. Similar sequences in the cp-genome of *S. miltiorrhiza* were isolated and used as the basis of genome assembly. Second, the cp-genome of *Sesamum indicum* was selected for guiding the order of contigs because of its highest similarity with the *S. miltiorrhiza* cp-genome. Third, to fill the gaps in the assembly, isolated sequences and contigs were used to repeatedly search against SMRT reads of *S. miltiorrhiza* gDNA. Then, an initial assembly was obtained by

TABLE 2 Chloroplast genome assemblies obtained using TGS techniques.

Family	Species	Technique	Total size/kb	Structure	Protein-coding genes	rRNA	tRNA	GC	Reference
Dicotyledonae									
Aristolochiaceae	*Asarum heterotropoides*	O,I	190.2	Tri	100	8	38	36.78%	(Yoo et al., 2021)
	Asarum maculatum	O,I	193.1	Tri	100	8	38	36.24%	(Yoo et al., 2021)
	Asarum misandrum	O,I	193.2	Tri	99	8	38	36.22%	(Yoo et al., 2021)
Asteraceae	*Carthamus tinctorius*	P	153.0	Qua	79 (Uni)	4 (Uni)	29 (Uni)	37.80%	(Wu et al., 2019)
	Chrysanthemum boreale	P	151.0	Qua	87	8	46	37.47%	(Won et al., 2018)
Cucurbitaceae	*Luffa acutangula*	P,I	157.2	Qua	87	8	36	37.14%	(Yundaeng et al., 2020)
	Luffa aegyptiaca	P,I	157.3	Qua	87	8	36	37.12%	(Yundaeng et al., 2020)
Fabaceae	*Callerya reticulata*	O,I	132.5	Bi	74	4	30	34.19%	(Chen et al., 2022b)
	Callerya nitida	O,I	132.4	Bi	74	4	30	33.89%	(Chen et al., 2022b)
Lamiaceae	*Salvia miltiorrhiza*	P,I	151.3	Qua	86	8	37	38%	(Qian et al., 2013)
Ranunculaceae	*Aconitum barbatum* var. *puberulum*	P,S	156.7	Qua	84	34	8	38.7%	(Chen et al., 2015)
Gentianaceae	*Swertia mussotii*	P	153.4	Qua	84	8	37	38.20%	(Xiang et al., 2016)
Loranthaceae	*Taxillus chinensis*	P,I	121.4	Qua	66	8	28	37.3%	(Li et al., 2017b)
Loranthaceae	*Taxillus sutchuenensis*	P,I	122.6	Qua	66	8	28	37.3%	(Li et al., 2017b)
Caricaceae	*Vasconcellea pubescens*	O,I	158.7	Qua	82	8	37	37%	(Lin et al., 2020)
Nelumbonaceae	*Nelumbo nucifera*	P,I	163.6	Qua	85	8	37	/	(Wu et al., 2014)
Monocotyledoneae									
Liliaceae	*Fritillaria hupehensis*	P	152.1	Qua	89	8	38	36.97%	(Li et al., 2014)
	Fritillaria taipaiensis	P	151.7	Qua	89	8	38	36.97%	(Li et al., 2014)
	Fritillaria cirrhosa	P	152.0	Qua	89	8	38	36.95%	(Li et al., 2014)
	Fritillaria unibracteata var. *wabuensis*	P	151.0	Qua	88	8	37	37.0%	(Li et al., 2016)
	Lilium rosthornii	P,I	152.2	Qua	85	8	38	37.02%	(Wu et al., 2021a)
Poaceae	*Coix lacryma-jobi*	P,I	140.9	Qua	87 (Uni)	4 (Uni)	32 (Uni)	/	(Kang et al., 2018)
Zingiberaceae	*Curcuma longa*	P,I	162.2	Qua	87	8	36	36.20%	(Li et al., 2019)
Araceae	*Spirodela polyrhiza*	P	169.0	Qua	85	8	36	35.68%	(Zhang et al., 2020b)
Orchidaceae	*Dendrobium officinale*	P,I	152.2	Qua	89	8	30	37.46%	(Zhong et al., 2016)

P, PacBio SMRT; O, Oxford Nanopore; I, Illumina; S, Sanger;/, not reported; Tri, Tripartite; Qua, Quadripartite; Bi, Bipartite; Uni, the number of unique genes.

TABLE 3 Mitochondrial genome assemblies obtained using TGS techniques.

Family	Species	Technique	Total size/kb	GC content	Protein-coding genes	tRNA	rRNA	Repeat sequence	Reference
Magnoliaceae	*Magnolia biondii*	O	967.1	46.6%	41	20	3	27%	(Dong et al., 2020)
Lamiaceae	*Scutellaria tsinyunensis* Conformation A	O,I,S	354.1	45.26%	32	24	3	/	(Li et al., 2021b)
	Scutellaria tsinyunensis Conformation B	O,I,S	255.7, 98.4	45.26%	32	24	3	/	(Li et al., 2021b)
Fabaceae	*Dalbergia odorifera*	P,I	435	45.1%	33	17	4	4.0%	(Hong et al., 2021)
Fabaceae	*Sophora japonica*	P,I	484.916	45.4%	32	17	3	4%	(Shi et al., 2018)
Umbelliferae	*Coriandrum sativum* circle 1	O,I	82.926	/	14	12	2	/	(Wang et al., 2021c)
Umbelliferae	*Coriandrum sativum* circle 2	O,I	224.59	/	41	16	5	/	(Wang et al., 2021c)

P, PacBio SMRT; O, Oxford Nanopore; I, Illumina; S, Sanger;/, not reported.

extending the contigs, adding new reads, and conducting reassembly. Finally, the regions of junction between IRs and LSC (or SSC) were amplified and sequenced by Sanger sequencing and the final assembly was obtained by integrating the Sanger sequences into the initial assembly using Seqman (DNASTAR, WI). Strand-specific RNA sequencing was also conducted to determine the expression level of genes in the cp-genome of *S. miltiorrhiza*.

Further data analysis: The RNA-seq reads were mapped to the final assembly of *S. miltiorrhiza* cp-genome using Tophat to identify polycistrons and non-coding RNA (ncRNA) and to determine the content of protein-coding transcripts (cRNA) and ncRNA. Strand-specific real-time quantitative PCR (ss-qPCR) was also conducted to validate the results of RNA-seq. DNA modifications were predicted using the SMRT Portal software (v1.3.2). As a result, the authors identified 19 polycistronic transcripts containing 71 genes, which consisted of 58 protein-coding genes, four rRNA, and nine tRNA. Furthermore, 136 ncRNA transcripts were identified and classified into two categories, intergenic ncRNA and antisense ncRNA (asRNA). Using SMRT Portal 1.3.2, two DNA modification motifs and 2687 DNA modification sites were predicted. Interactions between asRNA and cRNA, DNA modification and gene expression were also analyzed. The results showed that the expression level of protein-coding genes was positively associated with that of asRNA, and the DNA modification was correlated with higher expression of ncRNA.

In addition to the software mentioned above, many kinds of newly developed software, such as GetOrganelle (Jin et al., 2020), Fast-Plast (https://github.com/mrmckain/Fast-Plast), CPGAVAS2 (Shi et al., 2019), and CPGview (Liu et al., 2023), are also available for *de novo* assembly, annotation, analysis and visualization of chloroplast genomes.

The mitochondrial genomes of medicinal plants are more complex than cp-genomes, which usually contain multiple conformations (isoforms) instead of circular molecules (Kozik et al., 2019; Wang et al., 2021c). Previously, many mitochondrial genome assemblies failed to obtain all the isoforms of the mitochondrial genome of medicinal plants because of the limitations of the methods (Kozik et al., 2019). With the help of TGS techniques, researchers successfully captured various conformations of the mitochondrial genomes of *Coriandrum sativum* (Wang et al., 2021c), *Scutellaria tsinyunensis* (Li et al., 2021b), and several other valuable medicinal plants, providing more complete and precise references for the further utilization of mitochondrial genome sequences.

3.2 Revelation of transcriptomes

Although combining whole genome sequencing with transcriptomic analysis is a useful strategy for characterizing the genetic information of medicinal plants, it is too expensive to obtain enough TGS and short reads data for *de novo* genome assembly. Species with high-quality reference genomes only account for a small proportion of medicinal plants. Therefore, finding an ideal strategy for characterizing genetic information of medicinal plants without reference genomes becomes important and promising. The emergence of TGS allowed researchers to obtain full-length transcriptomes at an isoform level at a low cost. Due to the application of TGS, mainly ONT and SMRT, to RNA sequencing, the transcriptome analysis methods have gradually been revolutionized (Zhao et al., 2019a). To date, more than 25 transcriptomes of medicinal plants, such as *S. miltiorrhiza* (Xu et al., 2015; Xu et al., 2016b), *Dendrobium officinale* (He et al., 2017), *Drynaria roosii* (Sun et al., 2018b), *Astragalus membranaceus* (Li et al., 2017a), and many other species have been revealed using TGS. Herein, we choose *A. membranaceus* as an example to demonstrate the specific process of applying TGS to transcriptomic analysis of species without reference genomes (Li et al., 2017a).

Sample and library preparation: In this study, taproots and leaves from *A. membranaceus* were collected, washed, and then stored in liquid nitrogen as samples, followed by RNA extraction using Spectrum Plant

Total RNA Kit. The extracted RNAs were assessed using an Agilent 2100 Bioanalyzer, among which high-quality RNAs were utilized to prepare first-strand cDNA. Then, the first strand of cDNA was used to synthesize and amplify the second strand of cDNA. Finally, an Iso-seq library preparation was finished with 400 μl of cDNA from each sample.

Sequencing and data processing: The Iso-seq libraries were sequenced on the PacBio RSII with three SMRT cells for 1–2 kb libraries and five SMRT cells for 2–3 kb libraries. Assembly was not needed in this experiment. The raw data were processed with the standard RS_Iso-Seq protocol (SMRT Analysis 2.3). Specifically, according to the results of polyA tails and primers detection, 494,408 reads of inserts (ROIs) for leaf tissue and 500,007 ROIs for root tissue in the raw data were classified as full-length and non-full-length reads. The authors obtained 115,725 full-length consensus sequences for leaf tissue and 102,334 for root tissue from full-length ROIs and clustered them into different isoforms, followed by polishing with non-full-length ROIs. Full-length consensus sequences with more than 99% accuracy were classified as high-quality (HQ) transcripts, while other sequences were classified as low-quality (LQ) transcripts using Quiver. As a result, researchers generated 75,816 HQ transcripts and 39,909 LQ transcripts for leaf tissue and 73,755 HQ transcripts and 28579 LQ transcripts for root tissue. Finally, HQ and LQ transcripts were corrected with an Illumina RNA-seq paired-end data set followed by redundancy removal using the CD-HITv4.6 package.

Further data analysis: For isoform identification, the non-redundant transcripts were clustered into families using the Coding GENome reconstruction Tool (Cogent v1.4). Finally, these transcript families were reconstructed as one or more unique transcript models through the De Bruijn graph method. Mapping the non-redundant transcripts to the unique transcript models, splicing junctions for transcripts were examined. Transcription isoforms of unique transcript models were identified by collapsing transcripts with identical splicing junctions, and SUPPA was used to detect alternative splicing events.

Functional annotation: Four protein databases (UniProtKB_Viridiplantae, UniProtKB_MEDTR, UniProtKB_SOYBN, and the curated soybean reference protein annotation) were used for functional annotation of unique transcript models using BlASTX (NCBI-BLAST v2.2.27+) and unique transcript models were then classified using GO and KEGG based on the best hit from UniProtKB_SOYBN.

Long non-coding RNA (LncRNA) identification: After removing annotated transcripts and filtering out unique transcript models with ORFs with a length of more than 100 amino acids or 50 amino acids at the end(s) internally, LncRNAs were annotated using Coding Potential Calculator v0.9r2 to assess ORF-filtered unique transcripts models.

Multiple cutting-edge transcriptomic-analysis software have recently been developed, such as 3GOLD (Logan et al., 2022) and MeShClust v3.0 (Girgis, 2022) for high-speed or high-quality sequence clustering, RATTLE (de la Rubia et al., 2022) for reference-free reconstruction and quantification of transcripts and NanoSplicer (You et al., 2022) for identifying splice junctions.

The full-length transcriptomes obtained using TGS also provide valuable resources about the expression pattern and isoforms of many functional genes associated with the biosynthesis of active components in medicinal plants.

Alternative splicing events of muti-exon genes in multicellular eukaryotes can enhance the functional diversity of the encoded proteins and regulate gene expression through complex post-transcriptional mechanisms (Reddy et al., 2013). TGS, with its long read lengths, can deliver high yields of long, full-length RNA or cDNA, supporting the quantification of genes and complete transcriptome analysis at the isoform level, which is especially useful for species without a reference genome (Xu et al., 2015).

3.3 Molecular identification

Current methods were insufficient for the quality control of multiple herbal ingredients in traditional Chinese patent medicines. Combining TGS with DNA barcoding has made it possible to monitor the quality of traditional Chinese patent medicines effectively and affordably, as verified in the study of Yimu Wan (Jia et al., 2017) and Jiuwei Qianghuo Wan (Xin et al., 2018).

In this section, we describe the molecular identification of traditional Chinese patent medicine 'Yimu Wan' (YMW) as an example (Jia et al., 2017). In the selected study, two reference samples of YMW, RF01 and RF02, were used to establish a standard method for identification, which was then successfully applied to commercial YMW samples.

Sample preparation: The reference samples RF01 and RF02 were made in the laboratory under the guidance of the Chinese Pharmacopoeia. RF02 was formulated by weighing 10 g of the mixed powder of *Leonurus japonicas*, *Angelica sinensis*, *Ligusticum chuanxiong*, *Aucklandia lappa* and other recorded proportions. *Panax ginseng* powder was then added to one RF02 sample as a biological indicator. For RF01, only *P. ginseng* was spared. Finally, pills were molded by mixing these two samples with double-distilled water. One-hundred-twenty milligrams of the sample RF02 was used to isolate gDNA, the quality of obtained gDNA was assessed by Nanodrop 2000, and the DNA concentrations were determined using an Agilent 2100 bioanalyzer.

Library preparation: Before library construction, the gDNA underwent PCR and purification. For the PCR process, different primers were added to distinct samples to amplify ITS2 and *psbA-trnH*. And the universal ITS2 and *psbA-trnH* primers were ligated with two tags (5 bp) to differentiate the sequences from different regions. The PCR process was conducted as described in the Chinese Pharmacopoeia. After purification, the PCR products were used to construct a SMRT sequencing library using the SMRTbell Template Prep Kit 1.0.

Sequencing and data processing: ITS2 and *psbA-trnH* amplicon sequencing were conducted on the PacBio SMRT instrument. CCS sub-read datasets were obtained using SMRT Analysis Server 2.3.0 provided by PacBio. The CCS reads from RF02 were extracted according to the tags mentioned above, which were used to construct data libraries using Perl scripts. The CCS reads were clustered followed by removing redundant sequences, and then identified in the DNA Barcoding System for Identifying Herbal Medicine using BLAST.

Validating the standard method: To validate its replicability, the same procedure as the quality control protocol established above was conducted with RF01.

Applying the standard method to commercial YMW: Three batches of YMW produced by the same manufacturer were randomly bought from various drug stores. The same sample preparation and testing methods as RF01 and RF02 were used for these samples. As a result, this research successfully developed an effective protocol to assess the quality of traditional Chinese patent medicines using PacBio SMRT sequencing.

4 Advantages and challenges

4.1 Advantages

According to previous studies (Supplementary Table 1), nuclear genomes of medicinal plants are usually diploids or polyploids with large genome sizes, high heterozygosity, and high repeat sequences proportion. It is also demonstrated that the GC content of nuclear genomes of many medicinal plants is generally lower than 50% on average and unevenly distributed in different chromosomes (Shang et al., 2020; Sun et al., 2020; Wu et al., 2021b; Li et al., 2022a; Xu et al., 2022a). These features have brought traditional short reads methods under challenge. Specifically, large repetitive and high/low GC content regions principally account for the misassemblies and gaps in the final NGS genome assemblies (Salzberg and Yorke, 2005; Schmidt and Pearson, 2016; Guo et al., 2018). As for genomic variations, although precise detection of single nucleotide polymorphisms (SNPs) and indels can be achieved by NGS, structure variations (SVs) remain difficult to detect. Moreover, because the distance between variations exceeds the length of short reads, it is difficult for NGS techniques to link individual SNPs and indels together and phase haplotypes and alleles (van Dijk et al., 2018). However, these obstacles can be overcome by TGS. With long read lengths, TGS can span most of the repeat regions and large SVs in medicinal plant genomes. Genomic variations, including SNPs, indels, and SVs, are also naturally connected in the same long read, making it much easier to phase alleles or haplotypes (Stander et al., 2021). Several polypoid medicinal plants [such as *Triadica sebifera* (4n=88) (Luo et al., 2022), *Rehmannia glutinosa* (4n=56) (Ma et al., 2021), *Aquilegia oxysepala* var. *kansuensis* (4n=28) (Xie et al., 2020)], species with high genome heterozygosity [such as *Aloe vera* (11.3%) (Jaiswal et al., 2021), *Curcuma longa* (4.83%) (Chakraborty et al., 2021), *Gar. jasminoides* (2.2%) (Xu et al., 2020b)], and species with an extremely high proportion of repeat sequences [such as *Allium sativum* (91.3%) (Sun et al., 2020), *Panax notoginseng* (88.2%) (Yang et al., 2021b)], were all sequenced and assembled using SMRT, ONT, or both, yielding many high-quality chromosome-level assemblies.

Regarding transcriptome research, first, the vast majority of eukaryotic genes do not strictly conform to the 'one gene-one transcript' pattern. Instead, they often have several different isoforms. The application of TGS allows researchers to obtain full-length transcripts at an isoform level, even if a reference genome is not available (Li et al., 2017a). Second, with low GC bias, SMRT and ONT also allow more precise quantification of the expression level of genes than NGS techniques, facilitating research into expression patterns of important genes.

4.2 Challenges

First, when attempting nuclear genome sequencing of medicinal plants, it is difficult for ONT to achieve both high accuracy and extremely long read lengths when used alone. For example, when applied to the sequencing of the polyploid genome of *Veratrum dahuricum*, ONT produced ultra-long reads. However, by mapping the NGS reads against the ONT assemblies and a SMRT CCS assembly, researchers found that the coverage of three ONT assemblies ranged from 49.15% to 76.31%, much smaller than that of the SMRT CCS assembly (99.53%) (Zeng et al., 2022). A hybrid sequencing approach seems to be a good resolution because it has been shown in many medicinal organisms (Jain et al., 2018; Song et al., 2018; Wang et al., 2018) that combining NGS short reads with ONT can improve both the accuracy and completeness of obtained assemblies. However, this strategy also greatly increases the sequencing costs.

Second, when applied to chloroplast genome sequencing of medicinal plants, two most used sample preparation methods are isolating chloroplasts from the plant tissue (Li et al., 2014; Wu et al., 2014) and extracting chloroplast sequences from sequencing data of total DNA (Chen et al., 2014). However, the former method is difficult for most researchers who are not specialists in chloroplast extraction, while the latter method is expensive because it needs to sequence the whole genome of medicinal plants. For this reason, we consider that TGS is a poor choice for sequencing chloroplast genomes of medicinal plants.

5 Discussion

As this review has demonstrated, applying TGS, mainly SMRT and ONT, can greatly promote the development of herbal genomics. So far, the nuclear genomes of more than 100 medicinal plants have been sequenced using TGS, a large proportion of which were assembled to a chromosome level, while the organelle genomes of some important medicinal organisms have also been precisely assembled using TGS data. In addition, TGS is revolutionizing how transcriptomes of medicinal plants are analyzed by enabling the acquisition of full-length transcriptomes at an isoform level without a reference genome. Furthermore, TGS combined with DNA barcoding is also an effective and affordable approach to monitoring the compositions of traditional Chinese patent medicines. In a word, TGS has greatly contributed to herbal genomics and enriched the genetic information of organism-derived species. However, studies of molecular identification using TGS are still rare, making it a promising field to study further. Apart from the fields mentioned above, the epigenomics of medicinal organisms is also promising because TGS can directly detect methylations of DNA and RNA molecules. In addition, assembling gapless T2T genomes using PacBio HiFi reads and ONT ultra-long reads is a new trend in the genomic research of animals and parasites, and it greatly increased our understanding of telomere and centromere regions. T2T genomes are still rarely reported for medicinal organisms, which should be a focus of future work.

Author contributions

LG and JS designed the review. LG wrote the manuscript. WX, TX, and JS revised and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This review was funded by a grant from the National Natural Science Foundation of China (grant number, 82073966). This review was also funded by a grant from Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences (2022-I2M-2-001).

Acknowledgments

The authors wish to thank BioRender.com for providing tools for creating Figure 1 and Figure 2.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1124536/full#supplementary-material

References

Benjamini, Y., and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* 40 (10), e72. doi: 10.1093/nar/gks001

Castaño, C., Berlin, A., Brandström Durling, M., Ihrmark, K., Lindahl, B. D., Stenlid, J., et al. (2020). Optimized metabarcoding with pacific biosciences enables semi-quantitative analysis of fungal communities. *New Phytol.* 228 (3), 1149–1158. doi: 10.1111/nph.16731

Chakraborty, A., Mahajan, S., Jaiswal, S. K., and Sharma, V. K. (2021). Genome sequencing of turmeric provides evolutionary insights into its medicinal properties. *Commun. Biol.* 4 (1), 1193. doi: 10.1038/s42003-021-02720-y

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021a). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18 (2), 170–175. doi: 10.1038/s41592-020-01056-5

Cheng, Q. Q., Ouyang, Y., Tang, Z. Y., Lao, C. C., Zhang, Y. Y., Cheng, C. S., et al. (2021c). Review on the development and applications of medicinal plant genomes. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.791219

Chen, Z., Jin, C., Wang, X., Deng, Y., Tian, X., Li, X., et al. (2022b). Characterization of the complete chloroplast genome of four species in *Callerya*. *J. AOAC Int.* 106 (1), 146–155. doi: 10.1093/jaoacint/qsac097

Chen, X., Li, Q., Li, Y., Qian, J., and Han, J. (2015). Chloroplast genome of *Aconitum barbatum* var. *puberulum* (Ranunculaceae) derived from CCS reads using the PacBio RS platform. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00042

Chen, H., Zhang, J., Yuan, G., and Liu, C. (2014). Complex interplay among DNA modification, noncoding RNA expression and protein-coding RNA expression in *Salvia miltiorrhiza* chloroplast genome. *PLoS One* 9 (6), e99314. doi: 10.1371/journal.pone.0099314

Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol* 4 (4), 265–270. doi: 10.1038/nnano.2009.12

de la Rubia, I., Srivastava, A., Xue, W., Indi, J. A., Carbonell-Sala, S., Lagarde, J., et al. (2022). RATTLE: reference-free reconstruction and quantification of transcriptomes from nanopore sequencing. *Genome Biol.* 23 (1), 153. doi: 10.1186/s13059-022-02715-w

Deng, Y., Liu, S., Zhang, Y., Tan, J., Li, X., Chu, X., et al. (2022). A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. *Mol. Plant* 15 (8), 1268–1284. doi: 10.1016/j.molp.2022.06.010

Dong, S., Chen, L., Liu, Y., Wang, Y., Zhang, S., Yang, L., et al. (2020). The draft mitochondrial genome of *Magnolia biondii* and mitochondrial phylogenomics of angiosperms. *PLoS One* 15 (4), e0231020. doi: 10.1371/journal.pone.0231020

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323 (5910), 133–138. doi: 10.1126/science.1162986

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7 (6), 461–465. doi: 10.1038/nmeth.1459

Girgis, H. Z. (2022). MeShClust v3.0: High-quality clustering of DNA sequences using the mean shift algorithm and alignment-free identity scores. *BMC Genomics* 23 (1), 423. doi: 10.1186/s12864-022-08619-0

Guo, L., Winzer, T., Yang, X., Li, Y., Ning, Z., He, Z., et al. (2018). The opium poppy genome and morphinan production. *Science* 362 (6412), 343–347. doi: 10.1126/science.aat4096

Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., et al. (2008). Single-molecule DNA sequencing of a viral genome. *Science* 320 (5872), 106–109. doi: 10.1126/science.1150427

Holley, G., Beyter, D., Ingimundardottir, H., Møller, P. L., Kristmundsdottir, S., Eggertsson, H. P., et al. (2021). Ratatosk: Hybrid error correction of long reads enables accurate variant calling and assembly. *Genome Biol.* 22 (1), 28. doi: 10.1186/s13059-020-02244-4

Hong, Z., Liao, X., Ye, Y., Zhang, N., Yang, Z., Zhu, W., et al. (2021). A complete mitochondrial genome for fragrant Chinese rosewood (*Dalbergia odorifera*, fabaceae) with comparative analyses of genome structure and intergenomic sequence transfers. *BMC Genomics* 22 (1), 672. doi: 10.1186/s12864-021-07967-7

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36 (4), 338–345. doi: 10.1038/nbt.4060

Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). The Oxford nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17 (1), 239. doi: 10.1186/s13059-016-1103-0

Jain, M., Tyson, J. R., Loose, M., Ip, C. L. C., Eccles, D. A., O'Grady, J., et al. (2017). MinION analysis and reference consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Res* 6, 760. doi: 10.12688/f1000research.11354.1

Jaiswal, S. K., Mahajan, S., Chakraborty, A., Kumar, S., and Sharma, V. K. (2021). The genome sequence of *Aloe vera* reveals adaptive evolution of drought tolerance mechanisms. *iScience* 24 (2), 102079. doi: 10.1016/j.isci.2021.102079

Jia, J., Xu, Z., Xin, T., Shi, L., and Song, J. (2017). Quality control of the traditional patent medicine yimu wan based on SMRT sequencing and DNA barcoding. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00926

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., dePamphilis, C. W., Yi, T. S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 21 (1), 241. doi: 10.1186/s13059-020-02154-5

Kang, S. H., Lee, H. O., Shin, M. J., Kim, N. H., Choi, B. S., Kumar, M., et al. (2018). The complete chloroplast genome sequence of *Coix lacryma-jobi* l. (Poaceae), a cereal and medicinal crop. *Mitochondrial DNA B Resour* 3 (2), 980–981. doi: 10.1080/23802359.2018.1507653

Kozik, A., Rowan, B. A., Lavelle, D., Berke, L., Schranz, M. E., Michelmore, R. W., et al. (2019). The alternative reality of plant mitochondrial DNA: One ring does not rule them all. *PLoS Genet.* 15 (8), e1008373. doi: 10.1371/journal.pgen.1008373

LaBarre, B. A., Goncearenco, A., Petrykowska, H. M., Jaratlerdsiri, W., Bornman, M. S. R., Hayes, V. M., et al. (2019). MethylToSNP: identifying SNPs in illumina DNA methylation array data. *Epigenet. Chromatin* 12 (1), 79. doi: 10.1186/s13072-019-0321-6

Li, L. F., Cushman, S. A., He, Y. X., and Li, Y. (2020a). Genome sequencing and population genomics modeling provide insights into the local adaptation of weeping *forsythia*. *Hortic. Res.* 7, 130. doi: 10.1038/s41438-020-00352-7

Li, J., Harata-Lee, Y., Denton, M. D., Feng, Q., Rathjen, J. R., Qu, Z., et al. (2017a). Long read reference genome-free reconstruction of a full-length transcriptome from *Astragalus membranaceus* reveals transcript variants involved in bioactive compound biosynthesis. *Cell Discovery* 3, 17031. doi: 10.1038/celldisc.2017.31

Li, M., and Li, L. M. (2022). RegScaf: a regression approach to scaffolding. *Bioinformatics* 38 (10), 2675–2682. doi: 10.1093/bioinformatics/btac174

Li, Y., Li, Q., Li, X., Song, J., and Sun, C. (2016). Complete chloroplast genome sequence of *Fritillaria unibracteata* var. *wabuensis* based on SMRT sequencing technology. *Mitochondrial DNA A DNA Mapp Seq Anal.* 27 (5), 3757–3758. doi: 10.3109/19401736.2015.1079892

Li, Q., Li, Y., Song, J., Xu, H., Xu, J., Zhu, Y., et al. (2014). High-accuracy *de novo* assembly and SNP detection of chloroplast genomes using a SMRT circular consensus sequencing strategy. *New Phytol.* 204 (4), 1041–1049. doi: 10.1111/nph.12966

Li, J., Xu, Y., Shan, Y., Pei, X., Yong, S., Liu, C., et al. (2021b). Assembly of the complete mitochondrial genome of an endemic plant, *Scutellaria tsinyunensis*, revealed the existence of two conformations generated by a repeat-mediated recombination. *Planta* 254 (2), 36. doi: 10.1007/s00425-021-03684-3

Li, C. Y., Yang, L., Liu, Y., Xu, Z. G., Gao, J., Huang, Y. B., et al. (2022a). The sage genome provides insight into the evolutionary dynamics of diterpene biosynthesis gene cluster in plants. *Cell Rep.* 40 (7), 111236. doi: 10.1016/j.celrep.2022.111236

Li, D. M., Zhao, C. Y., and Xu, Y. C. (2019). Characterization and phylogenetic analysis of the complete chloroplast genome of *Curcuma longa* (Zingiberaceae). *Mitochondrial DNA B Resour* 4 (2), 2974–2975. doi: 10.1080/23802359.2019.1664343

Li, Y., Zhou, J. G., Chen, X. L., Cui, Y. X., Xu, Z. C., Li, Y. H., et al. (2017b). Gene losses and partial deletion of small single-copy regions of the chloroplast genomes of two hemiparasitic *Taxillus* species. *Sci. Rep.* 7 (1), 12834. doi: 10.1038/s41598-017-13401-4

Lin, Z., Zhou, P., Ma, X., Deng, Y., Liao, Z., Li, R., et al. (2020). Comparative analysis of chloroplast genomes in *Vasconcellea pubescens* A.DC. and *Carica papaya* l. *Sci. Rep.* 10 (1), 15799. doi: 10.1038/s41598-020-72769-y

Liu, S., Ni, Y., Li, J., Zhang, X., Yang, H., Chen, H., et al. (2023). CPGView: A package for visualizing detailed chloroplast genome structures. *Mol. Ecol. Resour.* doi: 10.1111/1755-0998.13729

Logan, R., Fleischmann, Z., Annis, S., Wehe, A. W., Tilly, J. L., Woods, D. C., et al. (2022). 3GOLD: optimized levenshtein distance for clustering third-generation sequencing data. *BMC Bioinf.* 23 (1), 95. doi: 10.1186/s12859-022-04637-7

Luo, J., Ren, W., Cai, G., Huang, L., Shen, X., Li, N., et al. (2022). The chromosome-scale genome sequence of *Triadica sebifera* provides insight into fatty acids and anthocyanin biosynthesis. *Commun. Biol.* 5 (1), 786. doi: 10.1038/s42003-022-03751-9

Ma, L., Dong, C., Song, C., Wang, X., Zheng, X., Niu, Y., et al. (2021). *De novo* genome assembly of the potent medicinal plant *Rehmannia glutinosa* using nanopore technology. *Comput. Struct. Biotechnol. J.* 19, 3954–3963. doi: 10.1016/j.csbj.2021.07.006

Manekar, S. C., and Sathe, S. R. (2018). A benchmark study of *k*-mer counting methods for high-throughput sequencing. *Gigascience* 7 (12), giy125. doi: 10.1093/gigascience/giy125

McIntyre, A. B. R., Rizzardi, L., Yu, A. M., Alexander, N., Rosen, G. L., Botkin, D. J., et al. (2016). Nanopore sequencing in microgravity. *NPJ Microgravity* 2, 16035. doi: 10.1038/npjmgrav.2016.35

Morisse, P., Marchet, C., Limasset, A., Lecroq, T., and Lefebvre, A. (2021). Scalable long read self-correction and assembly polishing with multiple sequence alignment. *Sci. Rep.* 11 (1), 761. doi: 10.1038/s41598-020-80757-5

Navrátilová, P., Toegelová, H., Tulpová, Z., Kuo, Y. T., Stein, N., Doležel, J., et al. (2022). Prospects of telomere-to-telomere assembly in barley: Analysis of sequence gaps in the MorexV3 reference genome. *Plant Biotechnol. J.* 20 (7), 1373–1386. doi: 10.1111/pbi.13816

Orlando, L., Ginolhac, A., Raghavan, M., Vilstrup, J., Rasmussen, M., Magnussen, K., et al. (2011). True single-molecule DNA sequencing of a pleistocene horse bone. *Genome Res.* 21 (10), 1705–1719. doi: 10.1101/gr.122747.111

Pockrandt, C., Steinegger, M., and Salzberg, S. L. (2022). PhyloCSF++: A fast and user-friendly implementation of PhyloCSF with annotation tools. *Bioinformatics* 38 (5), 1440–1442. doi: 10.1093/bioinformatics/btab756

Qian, J., Song, J., Gao, H., Zhu, Y., Xu, J., Pang, X., et al. (2013). The complete chloroplast genome sequence of the medicinal plant Salvia miltiorrhiza. *PLoS One* 8 (2), e57607. doi: 10.1371/journal.pone.0057607

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., et al. (2012). A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* 13, 341. doi: 10.1186/1471-2164-13-341

Reddy, A. S., Marquez, Y., Kalyna, M., and Barta, A. (2013). Complexity of the alternative splicing landscape in plants. *Plant Cell* 25 (10), 3657–3683. doi: 10.1105/tpc.113.117523

Riehl, K., Riccio, C., Miska, E. A., and Hemberg, M. (2022). TransposonUltimate: software for transposon classification, annotation and detection. *Nucleic Acids Res.* 50 (11), e64. doi: 10.1093/nar/gkac136

Roeh, S., Weber, P., Rex-Haffner, M., Deussing, J. M., Binder, E. B., and Jakovcevski, M. (2017). Sequencing on the SOLiD 5500xl system - in-depth characterization of the GC bias. *Nucleus* 8 (4), 370–380. doi: 10.1080/19491034.2017.1320461

Salzberg, S. L., and Yorke, J. A. (2005). Beware of mis-assembled genomes. *Bioinformatics* 21 (24), 4320–4321. doi: 10.1093/bioinformatics/bti769

Sato, M. P., Ogura, Y., Nakamura, K., Nishida, R., Gotoh, Y., Hayashi, M., et al. (2019). Comparison of the sequencing bias of currently available library preparation kits for illumina sequencing of bacterial genomes and metagenomes. *DNA Res.* 26 (5), 391–398. doi: 10.1093/dnares/dsz017

Schmidt, M. H. M., and Pearson, C. E. (2016). Disease-associated repeat instability and mismatch repair. *DNA Repair (Amst)* 38, 117–126. doi: 10.1016/j.dnarep.2015.11.008

Shang, J., Tian, J., Cheng, H., Yan, Q., Li, L., Jamal, A., et al. (2020). The chromosome-level wintersweet (*Chimonanthus praecox*) genome provides insights into floral scent biosynthesis and flowering in winter. *Genome Biol.* 21 (1), 200. doi: 10.1186/s13059-020-02088-y

Shen, G., Luo, Y., Yao, Y., Meng, G., Zhang, Y., Wang, Y., et al. (2022a). The discovery of a key prenyltransferase gene assisted by a chromosome-level *Epimedium pubescens* genome. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1034943

Shen, Q., Zhang, L., Liao, Z., Wang, S., Yan, T., Shi, P., et al. (2018). The genome of *Artemisia annua* provides insight into the evolution of asteraceae family and artemisinin biosynthesis. *Mol. Plant* 11 (6), 776–788. doi: 10.1016/j.molp.2018.03.015

Shi, L., Chen, H., Jiang, M., Wang, L., Wu, X., Huang, L., et al. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Res.* 47 (W1), W65–w73. doi: 10.1093/nar/gkz345

Shi, Y., Liu, Y., Zhang, S., Zou, R., Tang, J., Mu, W., et al. (2018). Assembly and comparative analysis of the complete mitochondrial genome sequence of *Sophora japonica* 'JinhuaiJ2'. *PloS One* 13 (8), e0202485. doi: 10.1371/journal.pone.0202485

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* 14 (4), 407–410. doi: 10.1038/nmeth.4184

Song, C., Liu, Y., Song, A., Dong, G., Zhao, H., Sun, W., et al. (2018). The *Chrysanthemum nankingense* genome provides insights into the evolution and diversification of *Chrysanthemum* flowers and medicinal traits. *Mol. Plant* 11 (12), 1482–1491. doi: 10.1016/j.molp.2018.10.003

Stander, E. A., Dugé de Bernonville, T., Papon, N., and Courdavault, V. (2021). Computational biotechnology guides elucidation of the biosynthesis of the plant anticancer drug camptothecin. *Comput. Struct. Biotechnol. J.* 19, 3659–3663. doi: 10.1016/j.csbj.2021.06.028

Sun, M. Y., Li, J. Y., Li, D., Huang, F. J., Wang, D., Li, H., et al. (2018b). Full-length transcriptome sequencing and modular organization analysis of the naringin/neoeriocitrin-related gene expression pattern in *Drynaria roosii*. *Plant Cell Physiol.* 59 (7), 1398–1414. doi: 10.1093/pcp/pcy072

Sun, X., Zhu, S., Li, N., Cheng, Y., Zhao, J., Qiao, X., et al. (2020). A chromosome-level genome assembly of garlic (*Allium sativum*) provides insights into genome evolution and allicin biosynthesis. *Mol. Plant* 13 (9), 1328–1339. doi: 10.1016/j.molp.2020.07.019

Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S., and Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* 38 (15), e159. doi: 10.1093/nar/gkq543

Tu, L., Su, P., Zhang, Z., Gao, L., Wang, J., Hu, T., et al. (2020). Genome of *Tripterygium wilfordii* and identification of cytochrome P450 involved in triptolide biosynthesis. *Nat. Commun.* 11 (1), 971. doi: 10.1038/s41467-020-14776-1

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The third revolution in sequencing technology. *Trends Genet.* 34 (9), 666–681. doi: 10.1016/j.tig.2018.05.008

Wang, M., Gu, Z., Fu, Z., and Jiang, D. (2021b). High-quality genome assembly of an important biodiesel plant, *Euphorbia lathyris* l. *DNA Res.* 28 (6), dsab022. doi: 10.1093/dnares/dsab022

Wang, Y., Lan, Q., Zhao, X., Wang, L., Yu, W., Wang, B., et al. (2021c). The complete mitochondrial genome of *Coriandrum sativum*. *Mitochondrial DNA B Resour* 6 (8), 2391–2392. doi: 10.1080/23802359.2021.1951131

Wang, W., Schalamun, M., Morales-Suarez, A., Kainer, D., Schwessinger, B., and Lanfear, R. (2018). Assembly of chloroplast genomes with long- and short-read data: a comparison of approaches using *Eucalyptus pauciflora* as a test case. *BMC Genomics* 19 (1), 977. doi: 10.1186/s12864-018-5348-8

Wang, B., Yang, X., Jia, Y., Xu, Y., Jia, P., Dang, N., et al. (2022a). High-quality *Arabidopsis thaliana*genome assembly with nanopore and HiFi longreads. *Genomics Proteomics Bioinf.* 20 (1), 4–13. doi: 10.1016/j.gpb.2021.08.003

Wei, C., Yang, H., Wang, S., Zhao, J., Liu, C., Gao, L., et al. (2018). Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality. *Proc. Natl. Acad. Sci. U.S.A.* 115 (18), E4151–e4158. doi: 10.1073/pnas.1719622115

Won, S. Y., Jung, J. A., and Kim, J. S. (2018). The complete chloroplast genome of *Chrysanthemum boreale* (Asteraceae). *Mitochondrial DNA B Resour* 3 (2), 549–550. doi: 10.1080/23802359.2018.1468225

Wu, H., Bai, W., Li, Z., He, S., Yuan, W., and Wu, J. (2021a). The complete chloroplast genome of *Lilium rosthornii* diels (Liliopsida: Liliaceae) from hunan, China. *Mitochondrial DNA B Resour* 6 (2), 553–554. doi: 10.1080/23802359.2021.1872452

Wu, Z., Gui, S., Quan, Z., Pan, L., Wang, S., Ke, W., et al. (2014). A precise chloroplast genome of *Nelumbo nucifera* (Nelumbonaceae) evaluated with Sanger, illumina MiSeq, and PacBio RS II sequencing platforms: insight into the plastid evolution of basal eudicots. *BMC Plant Biol.* 14, 289. doi: 10.1186/s12870-014-0289-0

Wu, Z. H., Liao, R., Dong, X., Qin, R., and Liu, H. (2019). Complete chloroplast genome sequence of *Carthamus tinctorius* l. from PacBio sequel platform. *Mitochondrial DNA B Resour* 4 (2), 2635–2636. doi: 10.1080/23802359.2019.1643799

Wu, Z., Liu, H., Zhan, W., Yu, Z., Qin, E., Liu, S., et al. (2021b). The chromosome-scale reference genome of safflower (*Carthamus tinctorius*) provides insights into linoleic acid and flavonoid biosynthesis. *Plant Biotechnol. J.* 19 (9), 1725–1742. doi: 10.1111/pbi.13586

Wu, H., Zhao, G., Gong, H., Li, J., Luo, C., He, X., et al. (2020). A high-quality sponge gourd (*Luffa cylindrica*) genome. *Hortic. Res.* 7 (1), 128. doi: 10.1038/s41438-020-00350-9

Xiang, B., Li, X., Qian, J., Wang, L., Ma, L., Tian, X., et al. (2016). The complete chloroplast genome sequence of the medicinal plant *Swertia mussotii* using the pacbio RS II platform. *Molecules* 21 (8), 1029. doi: 10.3390/molecules21081029

Xie, J., Zhao, H., Li, K., Zhang, R., Jiang, Y., Wang, M., et al. (2020). A chromosome-scale reference genome of *Aquilegia oxysepala* var. *kansuensis*. *Hortic. Res.* 7 (1), 113. doi: 10.1038/s41438-020-0328-y

Xin, T., Xu, Z., Jia, J., Leon, C., Hu, S., Lin, Y., et al. (2018). Biomonitoring for traditional herbal medicinal products using DNA metabarcoding and single molecule, real-time sequencing. *Acta Pharm. Sin. B* 8 (3), 488–497. doi: 10.1016/j.apsb.2017.10.001

Xin, T., Zhang, Y., Pu, X., Gao, R., Xu, Z., and Song, J. (2019). Trends in herbgenomics. *Sci. China Life Sci.* 62 (3), 288–308. doi: 10.1007/s11427-018-9352-7

Xu, Z., Luo, H., Ji, A., Zhang, X., Song, J., and Chen, S. (2016b). Global identification of the full-length transcripts and alternative splicing related to phenolic acid biosynthetic genes in *Salvia miltiorrhiza*. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00100

Xu, Q., Niu, S. C., Li, K. L., Zheng, P. J., Zhang, X. J., Jia, Y., et al. (2022b). Chromosome-scale assembly of the *Dendrobium nobile* genome provides insights into the molecular mechanism of the biosynthesis of the medicinal active ingredient of *Dendrobium*. *Front. Genet.* 13. doi: 10.3389/fgene.2022.844622

Xu, Z., Peters, R. J., Weirather, J., Luo, H., Liao, B., Zhang, X., et al. (2015). Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of *Salvia miltiorrhiza* and tanshinone biosynthesis. *Plant J.* 82 (6), 951–961. doi: 10.1111/tpj.12865

Xu, Z., Pu, X., Gao, R., Demurtas, O. C., Fleck, S. J., Richter, M., et al. (2020b). Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants. *BMC Biol.* 18 (1), 63. doi: 10.1186/s12915-020-00795-3

Xu, K. W., Wei, X. F., Lin, C. X., Zhang, M., Zhang, Q., Zhou, P., et al. (2022a). The chromosome-level holly (*Ilex latifolia*) genome reveals key enzymes in triterpenoid saponin biosynthesis and fruit color change. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.982323

Yang, Z., Chen, S., Wang, S., Hu, Y., Zhang, G., Dong, Y., et al. (2021b). Chromosomal-scale genome assembly of *Eleutherococcus senticosus* provides insights into chromosome evolution in araliaceae. *Mol. Ecol. Resour* 21 (7), 2204–2220. doi: 10.1111/1755-0998.13403

Yoo, M. J., Jin, D. P., Lee, H. O., and Lim, C. E. (2021). Complete plastome of three Korean *Asarum* (Aristolochiaceae): Confirmation tripartite structure within Korean *Asarum* and comparative analyses. *Plants (Basel)* 2056 10 (10), 2056. doi: 10.3390/plants10102056

You, Y., Clark, M. B., and Shim, H. (2022). NanoSplicer: accurate identification of splice junctions using Oxford nanopore sequencing. *Bioinformatics* 38 (15), 3741–3748. doi: 10.1093/bioinformatics/btac359

Yundaeng, C., Nawae, W., Naktang, C., Shearman, J. R., Sonthirod, C., Sangsrakru, D., et al. (2020). Chloroplast genome data of *Luffa acutangula* and *Luffa aegyptiaca* and their phylogenetic relationships. *Data Brief* 33, 106470. doi: 10.1016/j.dib.2020.106470

Zeng, P., Tian, Z., Han, Y., Zhang, W., Zhou, T., Peng, Y., et al. (2022). Comparison of ONT and CCS sequencing technologies on the polyploid genome of a medicinal plant showed that high error rate of ONT reads are not suitable for self-correction. *Chin. Med.* 17 (1), 94. doi: 10.1186/s13020-022-00644-1

Zhang, Y., An, D., Li, C., Zhao, Z., and Wang, W. (2020b). The complete chloroplast genome of greater duckweed (*Spirodela polyrhiza* 7498) using PacBio long reads: insights into the chloroplast evolution and transcription regulation. *BMC Genomics* 21 (1), 76. doi: 10.1186/s12864-020-6499-y

Zhao, L., Zhang, H., Kohnen, M. V., Prasad, K., Gu, L., and Reddy, A. S. N. (2019a). Analysis of transcriptome and epitranscriptome in plants using pacbio iso-seq and nanopore-based direct RNA sequencing. *Front. Genet.* 10. doi: 10.3389/fgene.2019.00253

Zhong, Z., Zhang, G., Lai, X., and Huang, S. (2016). The complete chloroplast genome sequence of a new variety of *Dendrobium officinale* 'zhong ke IV hao'. *Mitochondrial DNA B Resour* 1 (1), 669–670. doi: 10.1080/23802359.2016.1219632

Zhou, C., McCarthy, S. A., and Durbin, R. (2023). YaHS: yet another Hi-c scaffolding tool. *Bioinformatics* 39 (1), btac808. doi: 10.1093/bioinformatics/btac808