nature communications

Review article

https://doi.org/10.1038/s41467-025-57448-8

# A long road ahead to reliable and complete medicinal plant genomes

Received: 12 October 2024

Accepted: 24 February 2025

Published online: 04 March 2025

Check for updates

Ling-Tong Cheng[1], Zi-Long Wang[2], Qian-Hao Zhu[3], Min Ye[2] & Chu-Yu Ye[1]

Long-read DNA sequencing has propelled medicinal plant genomics forward, with over 400 genomes from 203 plants sequenced by February 2025. However, many genomes still have assembly and annotation flaws, with only 11 gapless telomere-to-telomere assemblies. The core challenge remains identifying genes linked to secondary metabolite biosynthesis, regulation and evolution. High-quality complete genomes are essential for characterizing biosynthetic gene clusters and for enabling robust functional genomics and synthetic biology applications. We propose to focus on achieving more complete genome assemblies in diverse varieties on the basis of refining the currently available ones, leverage lessons from crop genomics research, and apply the cutting-edge genomics technologies in research of medicinal plant genomics.

Plants have long been recognized for their therapeutic properties. Throughout history, civilizations used traditional herbal medicine to treat various diseases. Although much of this traditional knowledge has gradually given way to modern pharmaceuticals, plants continue to play a vital and irreplaceable role in human health today, serving as a powerful tool in the battle against diseases[1]. Medicinal plants, as sources of rich plant natural products (PNPs), are invaluable for new drug development and discovery of lead compounds[2–4]. The majority of new herbal drugs have been generated from the abundant secondary metabolites (such as alkaloids, terpenoids, and phenolic compounds) produced through plant metabolism. Many approved and clinical-trial drugs are derived from natural products, such as methylsalicylate (*Filipendula ulmaria*)[5], the morphine (*Papaver somniferum*)[6], the antigout agent colchicine (*Colchicum autumnale*)[7], digoxin and digitoxin (*Digitalis purpurea*) for cardiac disorder[8], tetrahydrocannabinol and cannabidiol (*Cannabis sativa*)[9], the anticancer agents vinblastine and vincristine (*Catharanthus roseus*)[10], and paclitaxel (*Taxus brevifolia*)[11]. Additionally, revelations from ethnobotanical records have led to the development of artemisinin from *Artemisia annua*, a powerful antimalarial drug[12].

A key challenge in harnessing PNPs is decoding their biosynthetic pathways to enable bioengineering and large-scale production through plant breeding or synthetic biology[13]. The lack of genomic data has long hampered efficient exploration of the metabolic pathways and associated genes/enzymes in high-value medicinal plants. Before the genomic era, characterizing these pathways was laborious and time-consuming, relying on isotope labeling, forward genetics, or sequence-homology-based gene cloning[14–16]. These methods often identified components of the pathways but were inefficient in resolving complete biosynthetic routes due to the complex nature of the enzymes involved. Although some progress has been made in identifying genes involved in the biosynthesis of compounds like vindoline, catharanthine[17], taxol[18], and morphine[19], the process was slow due to the involvement of extensive random testing[20].

Recent advancements in genome sequencing technologies, assembly algorithms, and annotation software, along with more powerful computing resources, have made it more feasible and cost-effective to obtain high-quality plant genomes[21]. As a result, many medicinal plants now have genome sequences available, serving as a foundation for identifying secondary metabolite pathways and understanding the networks regulating the pathways. Additionally, sequencing the genomes of medicinal plants aids in understanding their phylogenetic relationships, promoting breeding, and advancing plant-derived drug development.

[1]Institute of Crop Science, Zhejiang Key Laboratory of Crop Germplasm Innovation and Utilization, Zhejiang University, Hangzhou, China. [2]State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences, Peking University, Beijing, China. [3]CSIRO Agriculture and Food, Canberra, Australia. e-mail: yecy@zju.edu.cn

This Review provides an update on the current state of medicinal plant genome sequencing, highlighting certain flaws in genome assembly and annotation. We also analyze the challenges posed by heterozygosity, polyploidy, and repetitive sequences on deciphering medicinal plant genomes. We discuss the identification of key enzymes in biosynthesis pathways, the mapping of biosynthetic gene clusters (BGCs), and the evolutionary perspectives of specialized metabolite biosynthesis. Finally, we outline future research directions to advance medicinal plant studies, emphasizing the need for improved genome assemblies, leveraging lessons from crop genomics research, and applying cutting-edge genomic technologies.

## Current status of sequencing medicinal plant genomes

We defined medicinal plants based on the pharmacopoeias of eight countries (China, Egypt, Europe, India, Japan, Korea, Brazil, and the USA). Through extensive literature review and cross-referencing these databases of medicinal plant species and other databases, such as the TCMPG and N3 Plant Genome Database (http://ibi.zju.edu.cn/N3database/index.php), we found that, as of February 2025, genomes of a total of 431 medicinal plants across 203 species have been sequenced (Supplementary Data 1).

Taxonomically, the sequenced medicinal plant genomes are unevenly distributed across plant orders. Of the 137 described terrestrial plant orders, 48 have representative medicinal plants with sequenced genomes (Fig. 1). Some orders, such as *Lamiales*, *Asterales*, and *Fabales*, are particularly prolific, with high biodiversity of medicinally valuable plants and numerous sequenced species (Fig. 2)[22]. There is a striking disparity between the expected and observed numbers of genome assemblies across different orders of medicinal plants. Several well-known drug-rich orders, such as *Fabales* (e.g., *Glycyrrhiza glabra*), *Ranunculales* (e.g., *Coptis chinensis*), and *Apiales* (e.g., *Panax ginseng*), have a significant underrepresentation in genome assemblies relative to their species diversity (Fig. 2).

Advancements in sequencing technology have dramatically accelerated both the quality and quantity of medicinal plant genome assemblies. Over the past 20 years, the number of medicinal plant

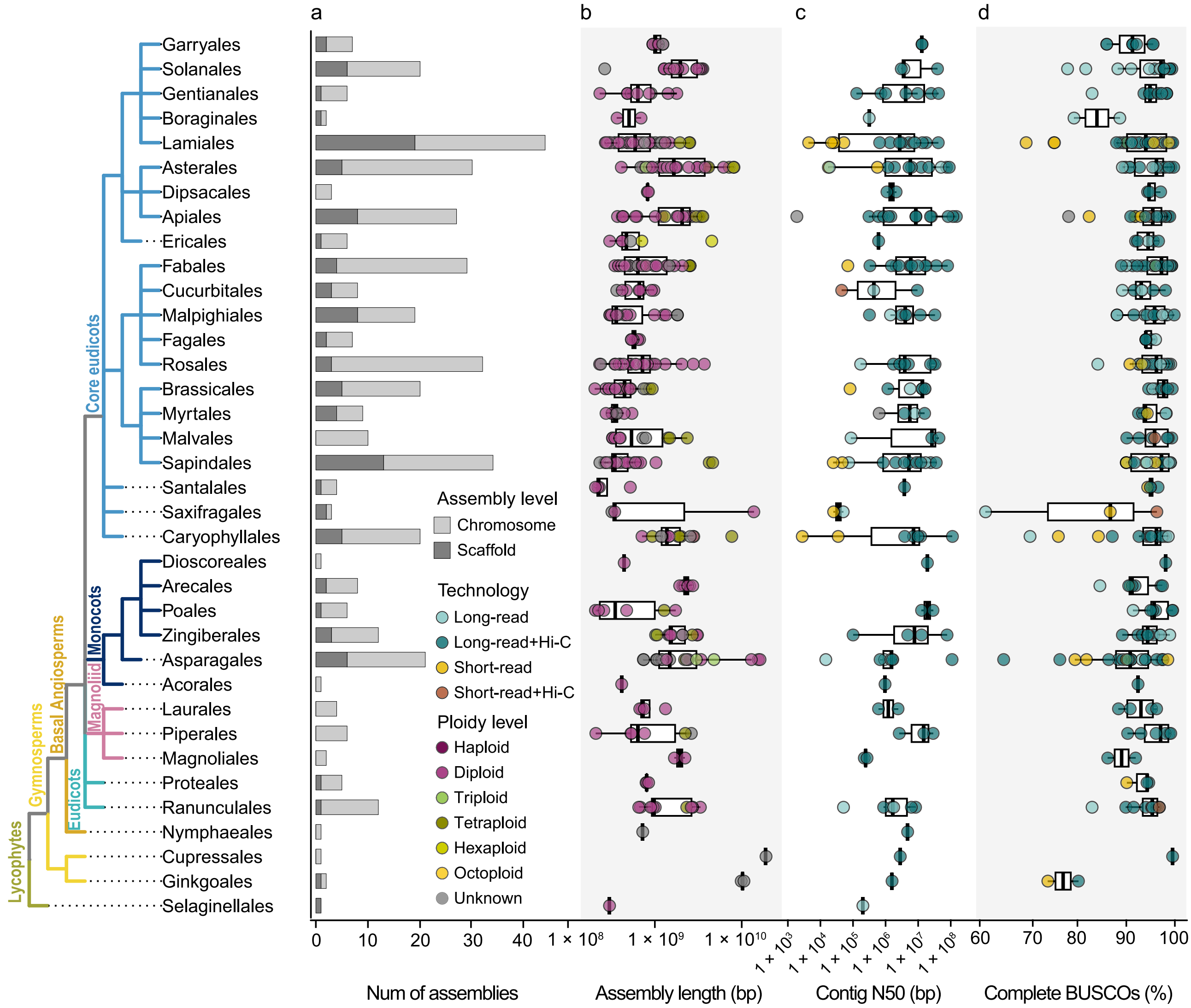

**Fig. 1 | Comparison of genome-size diversity and quality metrics across land plant orders with sequenced medicinal plant species. a** The numbers of genome assemblies at the chromosome and non-chromosome (scaffold) levels are shown for each lineage. **b** Box plots showing the distribution of assembly lengths of the medicinal plants in each plant order. **c** Box plots showing the distribution of contig N50 (the sequence length of the shortest contig at 50% of the total assembly size) for the medicinal plants in each plant order. **d** Box plots showing the distribution of complete BUSCO percentages of the published medicinal plant genomes in each plant order. Each dot represents a plant, which is color-coded based on ploidy level (**b**) or sequencing technology used (**c**, **d**). For all box plots, the box defines the interquartile range (25th–75th percentile), and the center line represents the median; whiskers extend to the maximum and minimum data values.

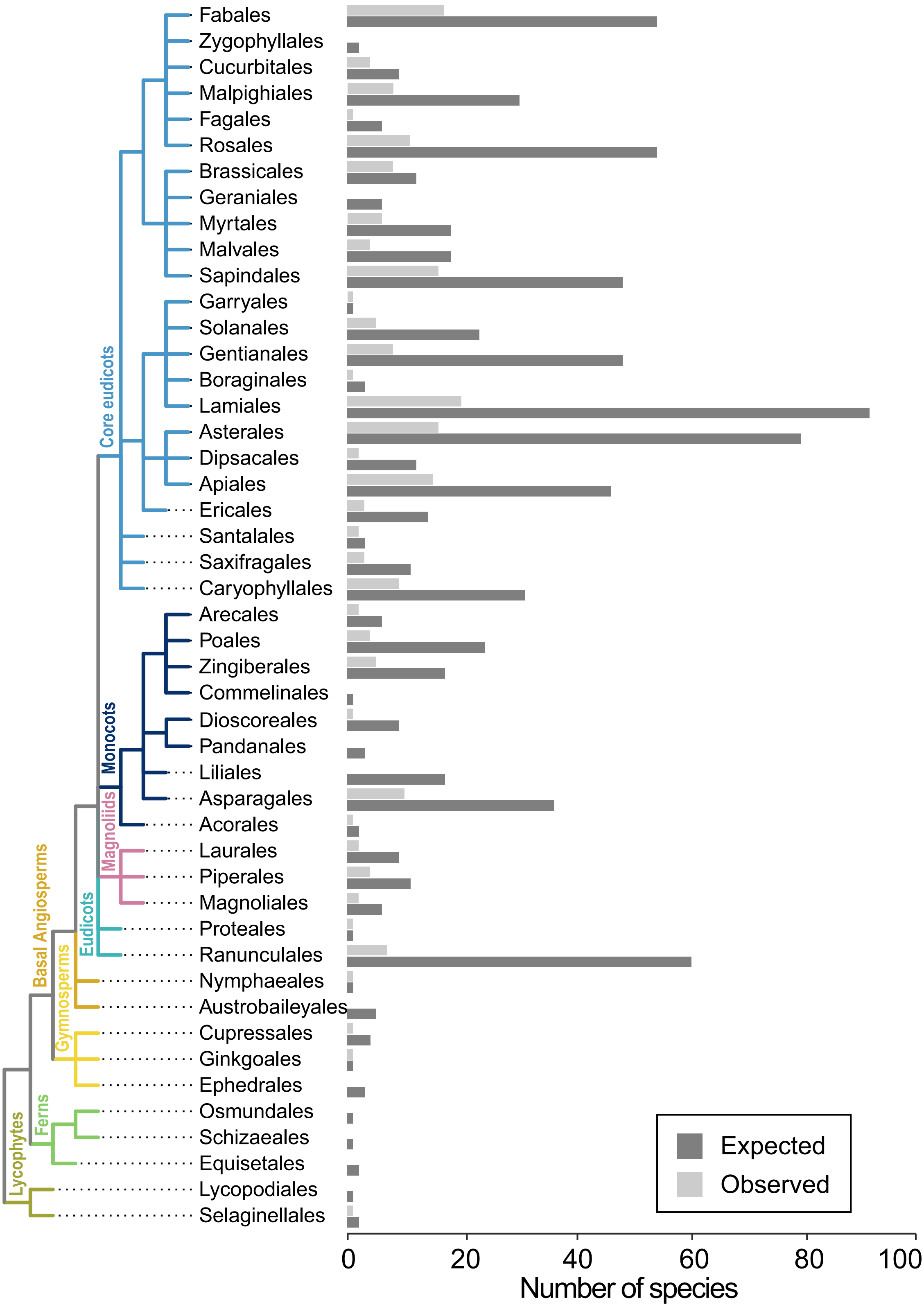

**Fig. 2 | Statistical representation of medicinal plant genome availability.** The number of species in each order with publicly available genome assemblies compared to the number expected based on species richness. The expected numbers (shown in dark gray) reflect the species richness within each order based on herb database, while the observed numbers (shown in light gray) indicate the actual genome assemblies available.

genome assemblies has grown exponentially, with 47.56% (205 assemblies) released in the past three years alone (Fig. 3a and Supplementary Data 1). The field is progressing rapidly and competitively, with some projects resulting in nearly simultaneous publications, such as two *Astragalus membranaceus* genomes[23,24] and two *Coptis chinensis* genomes[25,26]. The launch of initiatives such as the "1 K Medicinal Plant Genome Project" has driven the assembly of many high-quality genomes of Chinese herbal plants in recent years[27–31].

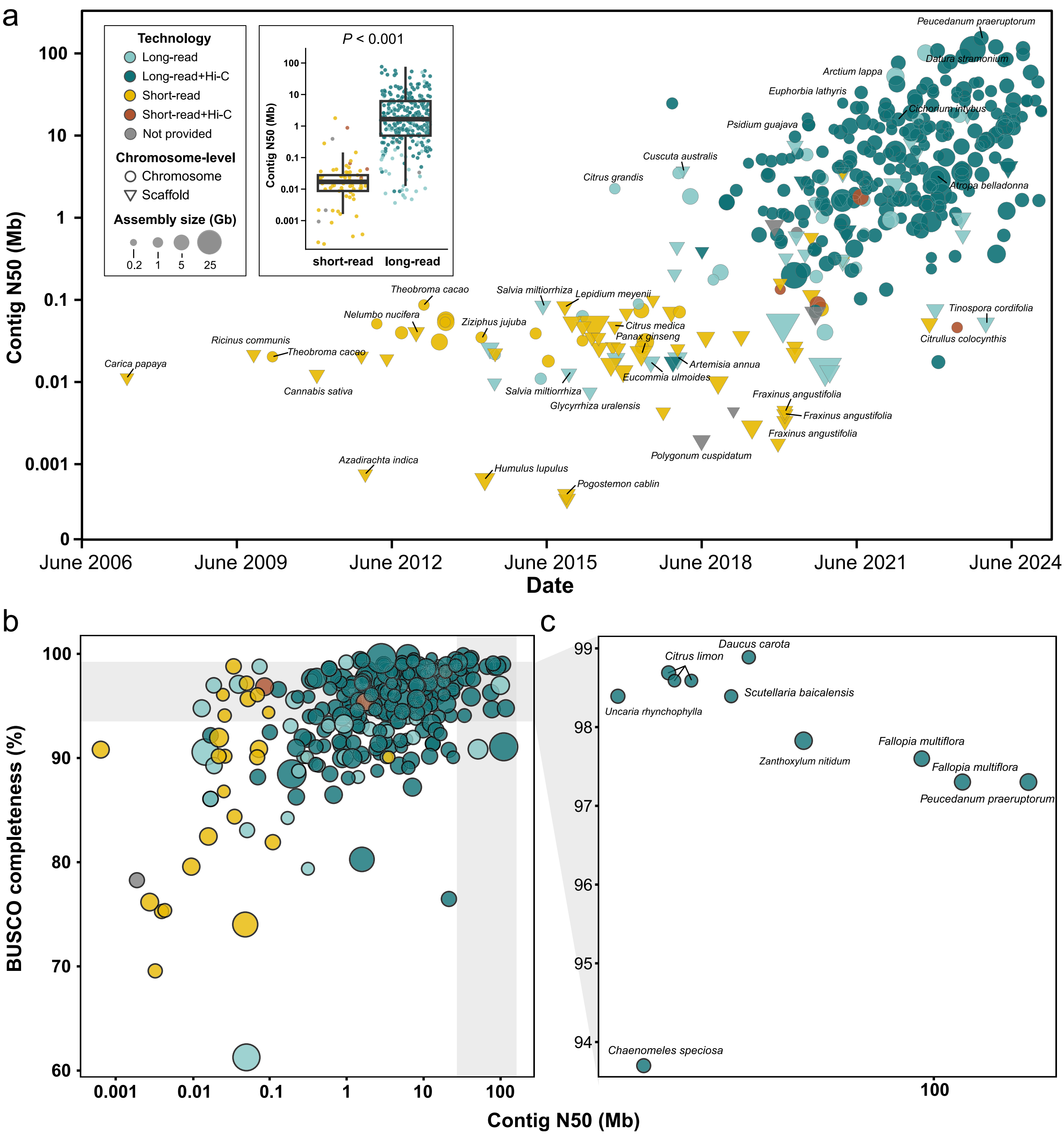

**Fig. 3 | Changes in quality and availability of medicinal plant genome assemblies over time. a** Assembly contiguity of the 487 medicinal plant species (plotted based on the release date and N50) with publicly available genome assemblies. Each dot or triangle represents a plant genome, which is colored based on the type of sequencing technology used and scaled by assembly size. The contiguity of the assemblies is significantly associated with the advent of long-read sequencing technology (Wilcoxon rank sum test, $p < 0.001$), alongside a marked increase in the annual number of genome assemblies. **b** Variation in contig N50 and BUSCO completeness across the available medicinal plant genomes. **c** The contig N50 and BUSCO completeness of the 26 T2T (telomere-to-telomere) assemblies.

Although the number of sequenced medicinal plant genomes has been steadily increasing, a significant portion (50.7%) remains in early assembly stages, with only a single version available (Supplementary Fig. 1), and 27 of these are still at the draft stage, severely limiting their usefulness. In many cases, there has been no effort to sequence additional varieties or update the sequenced assemblies, leaving them largely static and insufficient for broader research applications. Despite a small fraction of species that have been sequenced multiple times, either to improve assembly quality or to account for genetic variation across different varieties, the overall trend suggests that many projects have been content with simply having a genome and without further effort to refine the genomes or to cover more diversity.

In recent years, the prevalent sequencing strategy has shifted towards combining Illumina (second-generation sequencing or NGS) data with PacBio SMRT or ONT (third-generation sequencing or TGS) data. This approach is effective as it combines the strengths of both technologies. Illumina provides highly accurate short reads, while PacBio and ONT offer much longer reads that can span complex or repetitive regions of the genome to compensate the drawbacks of the short-read sequencing technologies. TGS technologies have dominated (98.04%) genome sequencing in the past three years. Over 304 nuclear genomes of medicinal plants have been sequenced using TGS, with 267 assembled to the chromosome level. Notably, 92.64% of the genomes assembled in the past three years have been mounted to chromosomes, largely due to the widespread (89.3%) adoption of chromosome conformation capture (Hi-C) techniques[32] and optical mapping[33], which capture the three-dimensional organization of chromosomes or provide long-range structural information to improve draft genome assemblies and yield chromosome-length scaffolds.

We summarized N50 and the percentage of Benchmarking Universal Single-Copy Orthologs (BUSCOs)[34] values to assess genome quality by estimating contig contiguity and completeness. N50, a commonly used contiguity measurement, has significantly increased with the advent of long-read sequencing, rising from 134.34 kb to 11.81 Mb (Fig. 3a). BUSCO is a widely used tool for assessing genome completeness by evaluating the presence of near-universal single-copy orthologs, which are highly conserved across different species. BUSCO completeness varies widely, ranging from 60% to 99% across the available genome assemblies (Fig. 3b), and N50 is not well-correlated with genome completeness. Numerous genomes exhibit high completeness across a broad range of N50 values, from high to low. Certain species like *Fraxinus angustifolia*, *Ginkgo bioba*, *Gastrodia elata*, and *Paeonia suffruticosa*, still have a low completeness despite relatively high contiguity (Fig. 3b). This is likely due to the inherent complexity and large size of their genomes, making it difficult to capture all regions accurately, even in highly contiguous assemblies. Additionally, no significant association was found between BUSCO percentages and genome size or ploidy level (Supplementary Fig. 2).

Recently, telomere-to-telomere (T2T) gapless assemblies have emerged as the gold standard for genome sequencing[35]. A T2T genome represents a high-quality, complete genome that includes all centromeres and repetitive regions, characterized by high accuracy, continuity, and integrity. To date, 11 genomes of medicinal plants have been assembled to T2T standards, including *Peucedanum praeruptorum*[36], *Scutellaria baicalensis*[37], *Fallopia multiflora*[38], and *Chaenomeles speciosa*[39]. These T2T assemblies exhibit a median contig N50 of 35.87 Mb (Fig. 3c) and an impressive completeness, with a BUSCO percentage ranging from 93.70% to 98.90% (Fig. 3c).

Besides the progress in sequencing technology, algorithms and toolkits used in de novo genome assembly have been improved, with a variety of them being employed across the sequenced medicinal plant genomes. Canu, Falcon, and Hifiasm were predominantly used in genome assembly, Pilon was the most used polishing tool, and LACHESIS and 3D-DNA were primarily used in scaffolding. Most assembly results were based on multiple software, and it is necessary to try different sorts of assembly software simultaneously. Furthermore, the choice of assembly tool might rely on the specific genomic characteristics being tackled (Supplementary Fig. 3). For example, SOAPdenovo2 and Platanus were frequently selected for assembling highly heterozygous genomes due to their ability to handle the complexity introduced by such variability. On the other hand, tools like Hifiasm and Falcon were preferred for genomes with high repeat content, leveraging their strength in resolving repetitive regions. The selection of assembly software is, therefore, a critical step in ensuring the accuracy and completeness of medicinal plant genome assemblies, with each tool offering distinct advantages depending on the genomic features of the species being studied.

To understand the worldwide contribution to sequencing medicinal plants, we identified the submitting institution for each genome assembly in our dataset and noticed a notable imbalance. China overwhelmingly leads de novo assembly efforts, contributing 69.9% (251 assemblies) of the total (Supplementary Fig. 4), followed by other Asian countries (38 assemblies), the USA (26 assemblies), and European nations (23 assemblies). Together, these regions contributed ~96.6% of all medicinal plant genome assemblies. These disparities likely stem from several factors, including varying levels of reliance on traditional medicine, the cultural and historical perspective, and the economic importance of medicinal plants.

## Evaluation of medicinal plant genome assemblies using key biosynthetic genes

An ideal genome should exhibit contiguity, completeness, and correctness, along with precise annotation. However, a significant gap persists between the ideal and the reality, which poses challenges in the practical application of medicinal plant research. Despite ongoing advancements in sequencing technology and efforts to enhance genome quality, achieving gap-free assemblies and error-free annotations continues to be elusive. The proportion of T2T assemblies is notably small, with most genomes remaining incomplete. Assessing the quality of a genome is challenging, particularly because it cannot rely solely on contiguity metrics. Common contiguity metrics, such as N50, may not accurately reflect the biological completeness or correctness of an assembly[40,41]. Alternatively, it is a general practice to assess completeness by using BUSCO or marker genes that are expected to be present[41], which evaluates genome assembly integrity by comparing conserved sequences.

To further assess the quality of accessible medicinal plant genomes, we examined key biosynthetic pathways and the associated genes crucial for the synthesis of bioactive compounds in nine medicinal plants (Fig. 4a and Supplementary Data 2). These reference genomes vary in quality, reflecting the current state of genomic resources in this field, from chromosome-level assemblies, such as *Papaver somniferum*, *Ginkgo biloba*, and *Taxus chinensis*, to scaffold-level assemblies like *Artemisia annua*, *Panax ginseng*, and *Rehmannia glutinosa*. The T2T *Scutellaria baicalensis* genome, with a BUSCO score of 98.40%, was also included. Using the coding sequences (CDS) of genes that have been cloned and experimentally validated in a certain plant, we aligned the sequences to the corresponding genome assembly and annotated CDS using BLAST (Supplementary Data 3) and revealed varying degrees of gene absence and incompleteness across the genomes (Fig. 4a). Many assembled genomes missed critical genes of the biosynthetic pathways. For example, in *Panax ginseng*, four key genes involved in ginsenoside biosynthesis (*PgHMGS*, *UGT1*, *UGTPg29*, and *UGTPg45*) were absent from both the annotated CDS and the assembly sequences. In *A. annua*, known for artemisinin biosynthesis, two genes (*AaSES* and *AaHMGS*) were missing from both the annotated CDS and the genome sequence. Likewise, the gene *PsTNMT*, which encodes an enzyme related to Alkaloids synthesis, could not be found in *Papaver somniferum* genome. Missing of these genes might be caused by misassembling of the genome.

Beyond incomplete genome assemblies, the number and exact structure of predicted genes (genome annotation) may also be incorrect in some cases, affecting the utility of the genomes. Even when genes were identified, some exhibited incomplete sequences. For instance, in *Ginkgo biloba*, eight genes crucial for the biosynthesis of ginkgolides, including *GbDXR* and *GbMYBF10*, were missing from the annotated CDS although present in the genome sequence, and *GbMADS9* showed low identity in both annotated CDS and genome sequence (Fig. 4a). In *Taxus chinensis*, while no genes were completely missing, several genes (*GGPPS*, *TAT*, and *DBTNBT*) exhibited low sequence identity. *Scutellaria baicalensis*, one of the few gap-free medicinal plant assemblies, still missed *SbCCD1* and showed low identity of *SbCHS-1* in the annotated gene sequences (Fig. 4a). The issues of missingness and low identity are related to the level of the corresponding assembly, highlighting the importance of pursuing high-quality assemblies.

We also incorporated a comparison between a chromosome-level[42] and a T2T genome[36] for *P. praeruptorum* to demonstrate how a well-assembled genome can illuminate BGCs. The newly assembled

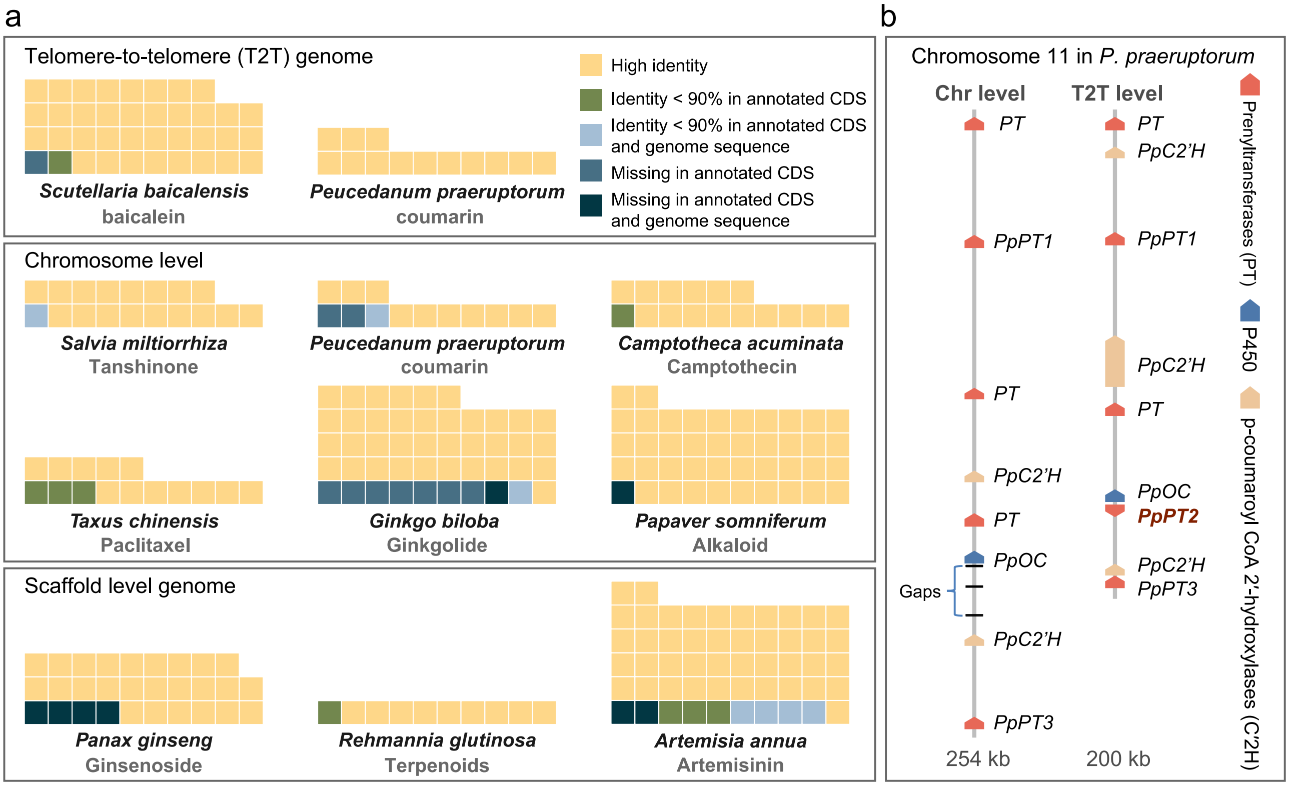

**Fig. 4 | Proportional representation of gene missingness and low identity in medicinal plant genomes. a** Gene gaps in medicinal plant genomes. For each species, the cloned genes (represented by color-coded rectangles) involved in synthesis of the bioactive secondary metabolite indicated below the species name were examined for their presence and absence as well as sequence identity in the corresponding species by BLAST search of the cloned gene against the CDS sequences and the full genomic sequences. A "miss" was defined as no hits at the e-value cutoff of 1e-50. "Identity <90%" represents cases where a hit was found, but the sequence identity between the cloned gene and the annotated CDS and/or the assembled genome sequence was below 90%. **b** Chromosomal mapping and identification of coumarin biosynthetic gene clusters in the chromosome-level (Chr level) and telomere-to-telomere (T2T) genomes of *Peucedanum praeruptorum*. Short horizontal lines indicate gaps in the genome, ~500 bp in length.

T2T genome shows significant improvements, with the contig N50 increasing from 11 Mb to 160 Mb and BUSCO scores rising from 96% to 98.2%. While the chromosome-level genome had an average of 27 contigs per chromosome, the T2T assembly consists of a single contig per chromosome, fully resolving telomeric and centromeric regions. Importantly, two key cytochrome P450 genes (*CYP71AJ49* and *CYP71AJ51*) involved in coumarin biosynthesis, initially missed in the chromosome-level genome, were correctly annotated in the T2T version (Fig. 4a). Moreover, key genes like *PpOC*, *PpPTs*, and *PpC'2H*, involved in coumarin biosynthesis, were found clustered on chromosome 11, forming a well-defined BGC. In contrast, in the chromosome-level assembly, the crucial gene *PpPT2* was incomplete due to three gaps in the region (Fig. 4b). As shown in this case, the T2T genome resolved these gaps, offering a complete genome and providing a clear roadmap (e.g., gene clusters, regulatory elements) for improving the production of the valuable compounds.

Although a limited number of genomes were used in the examination, the issues found reflect deficiencies in the completeness and correctness of assemblies and annotations of medicinal plant genomes. These issues could be due to problems during sequencing, assembly, or annotation processes, or they might stem from differences between the sequenced varieties and the ones from which the cloned genes were originally identified. No matter the causes, each step of genome assembly and gene annotation could introduce errors, affecting the final outputs. For most genome projects, genome assembly serves as a means to an end, neglecting that the ultimate goal of any genome project is to have a high-quality assembly for others to use. Hence, incomplete and sometimes erroneous genome assemblies impede downstream analysis and the design and interpretation of experiments. This underscores the need for ongoing improvements to address the errors and shadow in genome assemblies and annotations, thereby providing highly robust instruments for downstream genomic analyses.

## Factors impacting medicinal plant genomes de novo assembly

Plant genomes present unique challenges for high-quality sequencing and assembly due to a wide genome size range, polyploidy, genomic heterozygosity, and high repeat content. These obstacles are more pronounced in plants than in animals[43].

Heterozygosity refers to the presence of different alleles at a specific genetic locus in an individual's genome. It is commonly evaluated by analyzing the distribution of K-mer coverage and frequency of genome data. Heterozygosity complicates de novo sequencing and genome assembly. Among the sequenced medicinal plant genomes, 27.32% exhibited high heterozygosity (>1%), 49.76% moderate heterozygosity (0.5–1%), and only 22.92% fell within a manageable range (<0.5%) for assembly (Fig. 5a). Genome assembly typically relies on inbred, homozygous individuals treated as haploids, but medicinal plants often face challenges like self-incompatibility (e.g., *Echinacea purpurea*), frequent hybridization (e.g., *Panax ginseng*), and genetic diversity of wild accessions. While inbred lines could reduce heterozygosity, generating such lines is time-consuming and expensive. Additionally, many medicinal plants rely on heterozygosity and vegetative propagation to preserve valuable traits. To mitigate these challenges, strategies such as selecting varieties with low heterozygosity, inducing haploids, and employing haplotype-resolved assemblies have

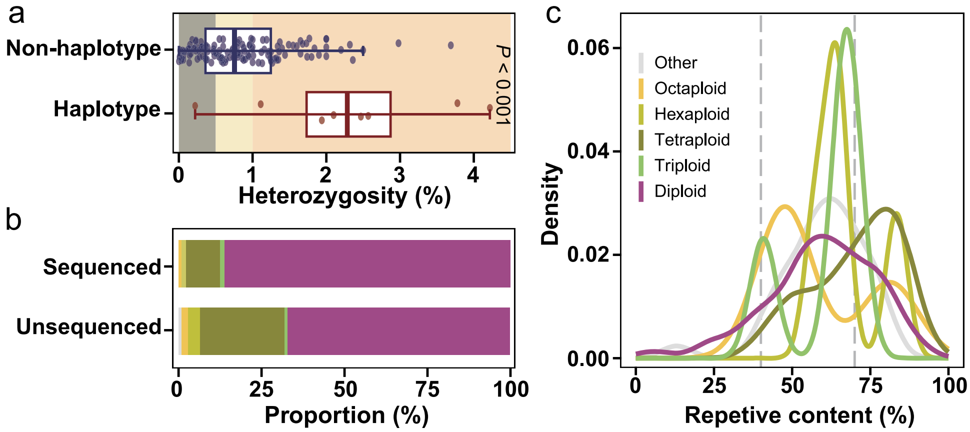

**Fig. 5 | Complex characteristics of medicinal plant genomes.** **a** Boxplot illustrating the heterozygosity of medicinal plants. Genomes assembled using haploid-resolved methods show significantly higher heterozygosity compared to non-haploid assemblies (Wilcoxon rank sum test, $p < 0.001$). **b** The proportion of polyploids in the sequenced and unsequenced medicinal plants. The ploidy of the sequenced plants was categorized based on genome survey results, while that of the unsequenced plants was estimated using the Kew Botanical Gardens Plant DNA C-values database[53]. Purple represents diploids, while green and yellow represent polyploids. The legend is the same as that in (**c**). **c** The distribution of repetitive sequence content across medicinal plants with different ploidy levels. The majority of publicly available genomes exhibit moderate (40–70%) to high (>70%) levels of repetitive sequences.

been used. For instance, the *Panax ginseng* line IR826 was chosen for sequencing due to its low heterozygosity[44]. Haploid materials were also used to simplify sequencing, as in *Eucommia ulmoides*[45] and *Ginkgo biloba*[46]. However, for many species, breeding homozygous varieties or collecting sufficient haploid tissue was impractical. In these cases, haplotype-resolved strategies offered another solution, with the 18 haplotype-resolved assemblies generated by the strategies showing significantly higher heterozygosity (2.80 ± 1.93%) compared to the non-haplotyped assembles (0.99 ± 1.19%) (Fig. 5a).

Polyploids pose significant challenges for genome assembly due to their genome complexity and size and sequence homology. Multiple closely related sub-genomes make it difficult to distinguish homoeologous loci, often leading to misassemblies, particularly in autopolyploids. We found that the majority of sequenced medicinal plant species were diploid (85.70%), with tetraploid species being the most common polyploids (10.10%) (Fig. 5b). Despite triploid and octoploid species like *Atropa belladonna* and *Phyllanthus emblica* have also recently been sequenced, sequencing of polyploids is still lagging behind that of diploids, because polyploid species is over represented in the unsequenced species, compared to the sequenced ones (Fig. 5b).

Complex eukaryotic genomes, including those of medicinal plants, contain a large number of repetitive sequences, which complicates the genome assembly process. Our analysis revealed that most (89.71%) publicly available medicinal plant genomes have over 40% repetitive content, with approximately one third (30.29%) exhibiting high levels (>70%) of repetitive sequences (Fig. 5c). This extensive repeat content often results in fragmented and incomplete genomes, with many unresolved regions caused by segmentally duplicated or other complex repeats. Therefore, accurately assembling and resolving repeat-rich genomes remains a significant challenge in current assembly efforts.

## Genomics-driven specialized metabolite biosynthesis and regulatory network analyses

Medicinal plants are highly valued for their active ingredients, often represented by specialized metabolites. A significant challenge in harnessing PNPs lies in deciphering the biosynthetic and regulatory mechanisms so as to enable bioengineering and large-scale production of PNPs through plant breeding or synthetic biology[47]. Before the genomic era, the absence of genomic information hindered the efficient exploration of the metabolic pathways and associated genes and enzymes in high-value medicinal herbs. Early efforts, such as isotope-labeled substrate feeding, struggled to resolve complex pathways[48]. With advances in molecular biology, the field shifted towards homology-based cloning and identifying mutants that altered metabolic profiles to isolate individual biosynthetic enzymes[15]. For instance, homologous gene cloning successfully identified key genes involved in the biosynthesis of taxol[49] and morphine[50].

Entering the twenty-first century, the advent of next-generation sequencing technologies and burgeoning synthetic biology tools ignited a renaissance of medicinal plant research. With the growing availability of genomic data for medicinal plants, it is now common to exploit high-throughput data mining combined with experimental validations to unravel the complex biosynthetic pathways underlying PNP accumulation[51]. Genomics-based strategies, including comparative genomics, functional genomics, and tissue-specific transcriptomics and metabolomics, were integrated to narrow down gene candidates. By comparing samples with contrasting levels of metabolite production, candidate genes can be identified based on their expression patterns. Candidates are validated through overexpression, CRISPR, or virus-induced gene silencing (VIGS), followed by metabolomic analysis to confirm their roles. This approach has been applied to identify genes involved in the biosynthesis of monoterpene indole alkaloids (MIAs) in *P. somniferum*[52,53], bisindole alkaloids in *C. roseus*[54], and polymethoxylated flavonoids in *C. reticulata*[55].

Furthermore, integrating genomics with metabolomics and rapid heterologous expression systems has provided a powerful tool to decode complex biosynthetic pathways[56]. Gene candidates can be expressed in microbial (e.g., *Escherichia coli*, *Saccharomyces cerevisiae*) or plant (e.g., *Nicotiana benthamiana*) systems to reconstruct pathways, followed by analyzing their metabolic profiles. This approach has been pivotal in discovering the pathways for the synthesis of hyoscyamine and scopolamine from nightshade plants[57], triptolide from *Tripterygium wilfordii*[58], colchicine from *Gloriosa superba*[59], etoposide lignans from mayapple[60], QS saponins (vaccine adjuvants) in soap bark tree *(Quillaja saponaria)*[61], hyperforin from *Hypericum perforatum*[62], and vinblastine in *C. roseus*[63]. Synthetic biology offers scalable solutions to the bottleneck in large-scale production of pharmacologically active PNPs. Genomics has greatly facilitated the decoding of entire pathways, enabling the de novo production of important compounds such as paclitaxel[64], vinblastine[65], protoberberine[66], astragaloside IV[67], breviscapine[68], and celastrol[69] in genetically engineered microbes or plant chassis.

In medicinal plants, specialized metabolites are often tightly regulated and subject to spatial, temporal, or condition-specific production, adding complexity to their dynamic regulation[70]. Environmental factors work together to affect active ingredient biosynthesis via hormone signal–transcription regulatory networks in medical plants. Transcription factors (TFs) from various families, such as bHLH, MYB, WRKY, AP2/ERF, bZIP, and NAC, are key regulators of these pathways and are often induced by hormones[71]. In the pre-genomic era, TF identification relied on promoter analysis and forward genetics. Genomic technologies have since revolutionized the identification and characterization of TFs involved in specialized metabolism. Whole-genome sequences allow for the systematic identification of cis-regulatory motifs (CRMs) in promoters, providing clues for potential TFs. For instance, the analysis of the jasmonic acid (JA)-responsive element in the *C. roseus ORCA3* promoter led to the discovery of TFs regulating monoterpene indole alkaloid (MIA) biosynthesis[72]. In *Bletilla striata*, the co-expression of *BsMYB2* with polysaccharide biosynthesis genes revealed its regulatory role[73]. Similarly, in *Salvia miltiorrhiza*, *SmbHLH60* was found to positively regulate phenolic acid and anthocyanin biosynthesis following methyl jasmonate (MeJA) treatment, as it showed an inverse expression pattern with key biosynthetic genes[74]. Another MeJA-responsive TF, *PgNAC72*, notably elevated total saponin levels by upregulating dammarenediol synthase (*PgDDS*) in *P. ginseng*[75]. Comparative genomics helps isolate homologous TFs across species. For example, *AaMYC2* in *A. annua* positively regulates artemisinin biosynthesis[76], while its homologs, *CrMYC2* in *C. roseus*[77] and *SmMYC2* in *S. miltiorrhiza*[78], regulate alkaloid and phenolic acid biosynthesis, respectively. Genomic tools have greatly advanced our understanding of transcriptional regulation in secondary metabolism.

## Mapping biosynthetic gene clusters in medicinal plants

A key breakthrough in plant genomics is the discovery of biosynthetic gene clusters (BGCs), which co-regulate specialized metabolite biosynthesis. Numerous BGCs have been identified in medicinal plants, driving secondary metabolite production (Supplementary Data 4). For instance, genome analysis of *Aesculus chinensis* revealed a gene cluster containing 4 oxidosqualene cyclases (*OSCs*) involved in triterpene saponin biosynthesis[79], and *Leonurus japonicus* was found to possess a cluster associated with leonurine alkaloid biosynthesis, which includes (UDP) glucosyltransferases (*UGTs*) and serine carboxypeptidase-like (*SCPL*) genes[80]. Similarly, in *Salvia officinalis*, clusters regulating diterpenoid production were identified, including genes encoding diterpene synthases (*diTPS*) and cytochrome P450s, showing organ-specific expression[81]. The genes of the entire biosynthesis pathway of glycyrrhizin were found to be clustered in the *Glycyrrhiza uralensis* genome[82].

Advances in genome mining and comparative genomics have accelerated the discovery of BGCs in medicinal plants. Tools like PlantiSMASH[83], Plant Cluster Finder[84], and PhytoClust[85] have facilitated the identification of these clusters. For instance, PlantiSMASH has been applied in *Taxus chinensis*, *Hypericum perforatum*, and *Salvia officinalis* to uncover clusters responsible for paclitaxel, hyperforin, and diterpenoid biosynthesis, respectively. These algorithms usually utilize synteny and co-expression patterns to identify potential metabolic gene clusters by recognizing signature and modification enzymes within the clusters. In *Taxus chinensis*, for example, the genes involved in paclitaxel biosynthesis (taxadiene synthase (*TS*) and cytochrome P450s) are organized in a cluster, exhibit root-specific expression patterns, and coordinately respond to jasmonate treatment[86].

The discovery of BGCs underscores the importance of high-quality genome assemblies. The proximity of genes in a BGC offers insights into pathway regulation, but incomplete or fragmented assemblies hinder BGC identification. For example, in *Catharanthus roseus*, a draft genome assembly was able to reveal 7 small clusters, each with two to three genes involved in vinblastine/vincristine biosynthesis, but its poor assembly quality (scaffold N50 size of 26–27 kb) made it difficult to resolve the relationships between these clusters[87]. Similarly, in *Salvia miltiorrhiza*, three key genes in tanshinone biosynthesis (*SmCPS1*, *SmCPS2*, and *SmKSL1*) were initially found to be clustered[88], but a more complete assembly later showed a cluster of five genes including the three mentioned above within a 310 kb region, and that all of the identified genes in this pathway except *CYP76AK1* are located in pseudochromosome 6[89]. As the advance in genome assembly and BGC mining technologies, more complete genomes will unlock deeper insights into plant metabolism.

## Evolutionary insights into biosynthesis of specialized metabolites

The diversity of medicinal compounds in plants is largely shaped by the sophisticated plant genome evolution. Certain metabolites are characteristics of closely related lineages, such as quinolizidine alkaloids in legumes, tropane and steroidal alkaloids in the Solanaceae family, and iridoids in labiates, implying a strong phylogenetic and ecological influence on their biosynthesis[90]. On the other hand, convergent evolution has also played a key role in the independent emergence of similar biosynthetic pathways across distantly related species, for example, the parallel evolution of cannabinoid formation in two genetically distant plants *Helichrysum umbraculigerum* and *Cannabis sativa*[91]. Moreover, there are notable exceptions where certain metabolites are absent (or present) in a specific taxon, even though neighboring and ancestral taxa exhibit the opposite trait. These deviations may reflect evolutionary shifts in key biosynthetic genes that emerged earlier in plant evolution, offering clues for identifying genes responsible for specific traits.

Phylogenetic analyses based on nuclear genes offer a more precise and less biased view compared to those based on ribosomal, mitochondrial, or chloroplast genes. With more plant genomes sequenced, lineage-wide characterization of biosynthetic genes becomes feasible. In the *Apiaceae* family, researchers used genomics and transcriptomics to explore why only certain species produce complex coumarins (CCs), aiming to elucidate the molecular mechanisms behind CC biosynthesis and diversification[92]. Similarly, in the genus *Salvia*, evolutionary analysis of 77 species has shown that diversification within the P450 subfamily CYP76AK contributed to changes in metabolite skeletons, linking enzyme functional divergence with chemical diversity[93]. The mint plant genome project has also offered insight into the distribution of volatile terpenoids in the *Lamiaceae* family and the diversification of the iridoid pathway across lineages[94,95].

Genomics-based phylogenetic studies could help track the origin and evolution of plant metabolites, shedding light on enzyme and metabolite evolution across different clades of the tree of life. For instance, benzylisoquinoline alkaloids (BIAs) are largely restricted to *Ranunculales* and *eumagnoliids*, and biochemical and molecular phylogenetic approaches have been employed to study BIA biosynthesis in basal angiosperms[96]. Similarly, research on the Nepeta lineage has uncovered mechanisms for the loss and re-evolution of iridoid biosynthesis, specifically the evolutionary origins of nepetalactone, a cat attractant found in catnip[97]. Another notable example is polymethoxyflavones (PMFs) in citrus, which possess potential anticancer properties. Comparative genomic and evolutionary analyses identified three key O-methyltransferase (OMT) genes involved in PMF biosynthesis, which are linked to varying PMF content in wild and cultivated mandarins[98].

Such evolutionary insights are crucial for discovering alternative resources and developing new medicinal compounds. By integrating evolutionary relationships into biosynthesis pipelines, researchers can more efficiently harness alternative or complementary species for therapeutic use. For instance, the close evolutionary relationship between *Aralia elata* and *Panax* ginseng has enabled researchers to manipulate the biosynthetic pathways of *A. elata* to produce valuable metabolites. Specifically, the partial deletion of the *DDS* gene in *A. elata* results in the absence of tetracyclic triterpenes, but overexpressing the *PgDDS* gene from *Panax ginseng* in *A. elata* callus restores saponin production[99], providing a powerful example of how evolutionary genomics can be leveraged to unlock the potential of medicinal plants.

## Outlook

**Towards complete and diverse genome assemblies.** The challenges ahead are formidable, but the continuing refinement of sequencing strategies and genomic analyses holds the promise of even greater discoveries in the future. Deciphering the synthesis, regulation, and evolution of secondary metabolites with invaluable pharmacological properties is a key challenge in medicinal plant genomics. The discovery and experimental validation of the genes responsible for these metabolites depend heavily on the availability of complete and high-quality genome assemblies. Lower-quality assemblies can lead to missing or misassembled genes, hindering accurate insights into metabolic pathways. Complete genome sequences, especially T2T genomes, help determine the organization of metabolic clusters.

Accurate and mature genome annotation is essential for facilitating downstream research and demands a combination of different sources of data, including ab initio gene prediction, spatiotemporal RNA-seq data, and full-length cDNA sequences. As mentioned above, some of the current issues on medicinal plant genomes stem from annotation, so reannotating already assembled genomes by incorporating more genomic data could offer a relatively quick solution to enhance the usability of the medicinal plant genomes.

In crops, a single reference genome is insufficient to capture full genomic diversity, leading to the pan-genome concept[100]. A pan-genome based on high-quality assemblies provides a comprehensive view of intraspecific diversity. In medicinal plants, intraspecific diversity is higher and to be linked to biosynthesis of specialized metabolites and adaptive traits. Understanding intraspecific genetic diversity can guide functional genomics of medicinal plants to improve the efficiency of drug discovery and can assist genomics-enabled breeding of designed medicinal varieties.

Genome sequences are crucial resources for the protection of endangered herbs and contribute greatly toward advancing breeding efforts. Many valuable or endangered medicinal plants remain under-represented in genomic research. For example, among the plants listed in the pharmacopeias of eight countries, 748 species (79.15%) have no reference genome. When considering the 4265 species in the Herb Database[101], the proportion rises to 89.38%. Prioritizing the future genomic research on these endangered and economically important medicinal plants will aid conservation and sustainable use.

**Leveraging the lessons from crop genomics research.** Though the focus of medicinal plant research differs, the well-established crop genomics approaches and concepts are transferable, including genetic diversity studies, phenotyping, and large-scale genome resequencing. Crop research places great emphasis on the systematic collection and phenotypic evaluation of germplasm resources, including both cultivated and wild species, as they form the foundation for genomic studies and breeding programs. Extensive germplasm collections help identify functional genes linked to complex traits, such as biosynthesis of pharmaceutical compounds and stress resilience, as to bridge the gap between genetic diversity, phenotypes, and the production of bioactive compounds. Phenomics platforms can support the precise collection of phenotypic data for dissecting traits like compound yield and stress tolerance in medicinal plants. Integrating phenomics with genomics can lead to the discovery of new genes or QTLs, aiding genomic selection of superior medicinal plant varieties.

Genome resequencing is key in identifying candidate genes linked to important traits in crops through genome-wide association studies (GWAS). While resequencing-based GWAS has been extensively applied to major crops like cereals and legumes, it remains largely underexplored in medicinal plants. To date, fewer than 1% of medicinal plant species have undergone resequencing. As more medicinal plant reference genomes become available, large-scale resequencing will enable gene discovery and understanding of domestication and evolution. For instance, the resequencing of 110 *Cannabis sativa* accessions has uncovered significant genetic differences between drug-type and hemp-type cultivars[102]. Safflower (*Carthamus tinctorius*) lines were resequenced in a GWAS to identify SNPs linked to the biosynthesis of bioactive compound hydroxysafflor yellow A[103].

**Harnessing new genomics technologies.** New genomics technologies, such as single-cell RNA sequencing and spatial transcriptomics, are revolutionizing our understanding of gene regulation in medicinal plants and the biosynthesis of specialized metabolites. Many key compounds are synthesized in distinct plant organs or cell types, driven by specialized gene expression patterns. For example, artemisinin is predominantly produced in the glandular trichomes of *A. annua*[104], tanshinones accumulate in the roots of *S. miltiorrhiza*[105], and taxol is derived from the bark of *Taxus*[106]. A clear example of how specific environmental and cultivation conditions influence metabolite production is the 'geoherb' effect[107], where the efficacy of medicinal plants is enhanced when grown under particular geographic and ecological settings. To untangle the intricate regulating network involving genetic and environmental components, future research should focus on integrating genomic resources with transcriptomic and proteomic data generated from different conditions.

Single-cell omics technologies allow study plant secondary metabolism with unprecedented resolution at the cellular level[108]. This innovation holds the promise of addressing long-standing questions about the uneven distribution and complex synthesis of specialized metabolites. Complementary single-cell multi-omics approaches have been successfully used to identify biosynthetic and regulatory pathways of secondary metabolites. In *C. roseus*, for example, a high-quality genome assembly identified gene clusters involved in monoterpenoid indole alkaloid (MIA) biosynthesis. Single-cell RNA sequencing showed MIA biosynthesis occurs sequentially in specific cell types within leaves. When combined with single-cell metabolomics, researchers were able to illustrate distinct patterns of metabolite localization and content across different cell types[109,110]. Similarly, single-cell analysis in *Hypericum perforatum* confirmed the hypothesis that hypericin biosynthesis depends on specific transparent glands in the leaves. This finding enabled the reconstruction of the entire hypericin biosynthesis pathway using yeast and tobacco systems, marking a major milestone in biosynthesis research driven by single-cell omics technologies[62]. These technological advances hold immense potential for identifying previously unknown genes involved in biosynthetic pathways and for revealing how these pathways are distributed across different cell types.

Looking forward, applying the new genomics technologies would offer a promising avenue for resolving the spatiotemporal gene networks regulating the biosynthesis of pharmaceutically important natural products and for their production through genetic engineering.

## References

1. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* **83**, 770–803 (2020).
2. Atanasov, A. G., Zotchev, S. B., Dirsch, V. M. & Supuran, C. T. Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.* **20**, 200–216 (2021).
3. Weng, J.-K., Philippe, R. N. & Noel, J. P. The rise of chemodiversity in plants. *Science* **336**, 1667–1670 (2012).

4. Li, F.-S. & Weng, J.-K. Demystifying traditional herbal medicine with modern approach. *Nat. Plants* **3**, 1–7 (2017).
5. Katanić, J. et al. In vitro and in vivo assessment of meadowsweet (Filipendula ulmaria) as anti-inflammatory agent. *J. Ethnopharmacol.* **193**, 627–636 (2016).
6. Labanca, F., Ovesna, J. & Milella, L. Papaver somniferum L. taxonomy, uses and new insight in poppy alkaloid pathways. *Phytochem. Rev.* **17**, 853–871 (2018).
7. Sapra, S. et al. Colchicine and its various physicochemical and biological aspects. *Med. Chem. Res.* **22**, 531–547 (2013).
8. Gurel, E., Karvar, S., Yucesan, B., Eker, I. & Sameeullah, M. An overview of cardenolides in digitalis-more than a cardiotonic compound. *Curr. Pharm. Des.* **23**, 5104–5114 (2017).
9. Zuardi, A. W., Crippa, J. Ad. S., Hallak, J. E. C., Moreira, F. & Guimarães, F. S. Cannabidiol, a Cannabis sativa constituent, as an antipsychotic drug. *Braz. J. Med. Biol. Res.* **39**, 421–429 (2006).
10. Neuss, N., Gorman, M., Svoboda, G., Maciak, G. & Beer, C. Vinca alkaloids. iii. 1 characterization of leurosine and vincaleukoblastine, new alkaloids from vinca rosea linn. *J. Am. Chem. Soc.* **81**, 4754–4755 (1959).
11. Wani, M. C., Taylor, H. L., Wall, M. E., Coggon, P. & McPhail, A. T. Plant antitumor agents. VI. Isolation and structure of taxol, a novel antileukemic and antitumor agent from Taxus brevifolia. *J. Am. Chem. Soc.* **93**, 2325–2327 (1971).
12. Tu, Y. The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine. *Nat. Med.* **17**, 1217–1220 (2011).
13. Chen, S. Biosynthesis of natural products from medicinal plants: challenges, progress and prospects. *Chin. Herb. Med.* **16**, 1–2 (2024).
14. Guo, L. et al. Natural products of medicinal plants: biosynthesis and bioengineering in post-genomic era. *Hortic. Res.* **9**, uhac223 (2022).
15. Bohlmann, J., Meyer-Gauen, G. & Croteau, R. Plant terpenoid synthases: molecular biology and phylogenetic analysis. *Proc. Natl. Acad. Sci. USA* **95**, 4126–4133 (1998).
16. Attieh, J. et al. Cloning and functional expression of two plant thiol methyltransferases: a new class of enzymes involved in the biosynthesis of sulfur volatiles. *Plant Mol. Biol.* **50**, 511–521 (2002).
17. Gao, J. et al. Biosynthesis of catharanthine in engineered Pichia pastoris. *Nat. Synth.* **2**, 231–242 (2023).
18. Croteau, R., Ketchum, R. E., Long, R. M., Kaspera, R. & Wildung, M. R. Taxol biosynthesis and molecular genetics. *Phytochem. Rev.* **5**, 75–97 (2006).
19. Ounaroon, A., Decker, G., Schmidt, J., Lottspeich, F. & Kutchan, T. M. (R, S)-Reticuline 7-O-methyltransferase and (R, S)-norcoclaurine 6-O-methyltransferase of Papaver somniferum–cDNA cloning and characterization of methyl transfer enzymes of alkaloid biosynthesis in opium poppy. *Plant J.* **36**, 808–819 (2003).
20. De Luca, V., Salim, V., Levac, D., Atsumi, S. M., Yu, F. Discovery and functional analysis of monoterpenoid indole alkaloid pathways in plants. In: *Methods in enzymology* (Elsevier 2012).
21. Xie, L. et al. Technology-enabled great leap in deciphering plant genomes. *Nat. Plants* **10**, 551–566 (2024).
22. Raja, R. R. Medicinally potential plants of labiatae (Lamiaceae) family: an. *Res. J. Med. Plant* **6**, 203–213 (2012).
23. Chen, Y. et al. A reference-grade genome assembly for Astragalus mongholicus and insights into the biosynthesis and high accumulation of triterpenoids and flavonoids in its roots. *Plant Commun.* **4**, 100469 (2023).
24. Fan, H. et al. Chromosome-scale genome assembly of Astragalus membranaceus using PacBio and Hi-C technologies. *Sci. Data* **11**, 1071 (2024).
25. Chen, D.-x et al. The chromosome-level reference genome of Coptis chinensis provides insights into genomic evolution and berberine biosynthesis. *Hortic. Res.* **8**, 121 (2021).
26. Liu, Y. et al. Analysis of the Coptis chinensis genome reveals the diversification of protoberberine-type alkaloids. *Nat. Commun.* **12**, 3276 (2021).
27. Chen, S.-L. et al. Molecular genetics research of medicinal plants. *China J. Chin. Mater. Med.* **44**, 2421–2432 (2019).
28. Zhang, G. et al. Hybrid de novo genome assembly of the Chinese herbal plant danshen (Salvia miltiorrhiza Bunge). *Gigascience* **4**, 62 (2015).
29. Sun, W. et al. The genome of the medicinal plant Andrographis paniculata provides insight into the biosynthesis of the bioactive diterpenoid neoandrographolide. *Plant J.* **97**, 841–857 (2019).
30. Zhang, J. et al. Genome of plant maca (Lepidium meyenii) illuminates genomic basis for high-altitude adaptation in the central Andes. *Mol. Plant* **9**, 1066–1077 (2016).
31. Chen, W. et al. Whole-genome sequencing and analysis of the Chinese herbal plant Panax notoginseng. *Mol. Plant* **10**, 899–902 (2017).
32. Belton, J.-M. et al. Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
33. Yuan, Y., Chung, C. Y.-L. & Chan, T.-F. Advances in optical mapping for genomic research. *Comput. Struct. Biotechnol. J.* **18**, 2051–2062 (2020).
34. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
35. Song, J.-M. et al. Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol. Plant* **14**, 1757–1767 (2021).
36. Bai, M. et al. The telomere-to-telomere (T2T) genome of Peucedanum praeruptorum Dunn provides insights into the genome evolution and coumarin biosynthesis. *GigaScience* **13**, giae025 (2024).
37. Pei, T. et al. Gap-free genome assembly and CYP450 gene family analysis reveal the biosynthesis of anthocyanins in Scutellaria baicalensis. *Hortic. Res.* **10**, uhad235 (2023).
38. Zeng, S. et al. T2T genome assemblies of Fallopia multiflora (Heshouwu) and F. multiflora var. angulata. *Sci. Data* **11**, 1103 (2024).
39. He, S. et al. A telomere-to-telomere reference genome provides genetic insight into the pentacyclic triterpenoid biosynthesis in Chaenomeles speciosa. *Hortic. Res.* **10**, uhad183 (2023).
40. Thrash, A., Hoffmann, F. & Perkins, A. Toward a more holistic method of genome assembly assessment. *BMC Bioinforma.* **21**, 249 (2020).
41. Jauhal, A. A. & Newcomb, R. D. Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. *Mol. Ecol. Resour.* **21**, 1416–1421 (2021).
42. Song, C. et al. A chromosome-scale genome of Peucedanum praeruptorum provide insights into Apioideae evolution and medicinal ingredient biosynthesis. *Int. J. Biol. Macromol.* **255**, 128218 (2024).
43. Sun, Y., Shang, L., Zhu, Q.-H., Fan, L. & Guo, L. Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci.* **27**, 391–401 (2022).
44. Xu, J. et al. Panax ginseng genome examination for ginsenoside biosynthesis. *Gigascience* **6**, gix093 (2017).
45. Li, Y. et al. High-quality de novo assembly of the Eucommia ulmoides haploid genome provides new insights into evolution and rubber biosynthesis. *Hortic. Res.* **7**, 183 (2020).
46. Guan, R. et al. Draft genome of the living fossil Ginkgo biloba. *Gigascience* **5**, 49 (2016).
47. Xin, T. et al. Trends in herbgenomics. *Sci. China Life Sci.* **62**, 288–308 (2019).

48. Herbert, R. B. The biosynthesis of plant alkaloids and nitrogenous microbial metabolites. *Nat. Prod. Rep.* **20**, 494–508 (2003).
49. Walker, K. & Croteau, R. Taxol biosynthetic genes. *Phytochemistry* **58**, 1–7 (2001).
50. Unterlinner, B., Lenz, R. & Kutchan, T. M. Molecular cloning and functional expression of codeinone reductase: the penultimate enzyme in morphine biosynthesis in the opium poppy Papaver somniferum. *Plant J.* **18**, 465–475 (1999).
51. Sun, W., Xu, Z., Song, C. & Chen, S. Herbgenomics: decipher molecular genetics of medicinal plants. *Innovation* **3**, 100322 (2022).
52. Wijekoon, C. P. & Facchini, P. J. Systematic knockdown of morphine pathway enzymes in opium poppy using virus-induced gene silencing. *Plant J.* **69**, 1052–1063 (2012).
53. Salim, V., Yu, F., Altarejos, J. & De Luca, V. Virus-induced gene silencing identifies C atharanthus roseus 7-deoxyloganic acid-7-hydroxylase, a step in iridoid and monoterpene indole alkaloid biosynthesis. *Plant J.* **76**, 754–765 (2013).
54. Liscombe, D. K. & O'Connor, S. E. A virus-induced gene silencing approach to understanding alkaloid metabolism in Catharanthus roseus. *Phytochemistry* **72**, 1969–1977 (2011).
55. Wen, J. et al. An integrated multi-omics approach reveals polymethoxylated flavonoid biosynthesis in Citrus reticulata cv. Chachiensis. *Nat. Commun.* **15**, 3991 (2024).
56. Dixon, R. A. & Dickinson, A. J. A century of studying plant secondary metabolism—from "what?" to "where, how, and why? *Plant Physiol.* **195**, 48–66 (2024).
57. Srinivasan, P. & Smolke, C. D. Biosynthesis of medicinal tropane alkaloids in yeast. *Nature* **585**, 614–619 (2020).
58. Zhang, Y. et al. Tandemly duplicated CYP82Ds catalyze 14-hydroxylation in triptolide biosynthesis and precursor production in Saccharomyces cerevisiae. *Nat. Commun.* **14**, 875 (2023).
59. Nett, R. S., Lau, W. & Sattely, E. S. Discovery and engineering of colchicine alkaloid biosynthesis. *Nature* **584**, 148–153 (2020).
60. Lau, W. & Sattely, E. S. Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone. *Science* **349**, 1224–1228 (2015).
61. Reed, J. et al. Elucidation of the pathway for biosynthesis of saponin adjuvants from the soapbark tree. *Science* **379**, 1252–1264 (2023).
62. Wu, S., Morotti, A. L. M., Yang, J., Wang, E. & Tatsis, E. C. Single-cell RNA sequencing facilitates the elucidation of the complete biosynthesis of the antidepressant hyperforin in St. John's wort. *Mol. Plant* **17**, 1439–1457 (2024).
63. Grzech, D., Hong, B., Caputi, L., Sonawane, P. D. & O'Connor, S. E. Engineering the biosynthesis of late-stage vinblastine precursors precondylocarpine acetate, catharanthine, tabersonine in Nicotiana benthamiana. *ACS Synth. Biol.* **12**, 27–34 (2022).
64. Jiang, B. et al. Characterization and heterologous reconstitution of Taxus biosynthetic enzymes leading to baccatin III. *Science* **383**, 622–629 (2024).
65. Caputi, L. et al. Missing enzymes in the biosynthesis of the anticancer drug vinblastine in Madagascar periwinkle. *Science* **360**, 1235–1239 (2018).
66. Jiao, X. et al. De novo production of protoberberine and benzophenanthridine alkaloids through metabolic engineering of yeast. *Nat. Commun.* **15**, 8759 (2024).
67. Xu, B. et al. Total biosynthesis of the medicinal triterpenoid saponin astragalosides. *Nat. Plants* **10**, 1826–1837 (2024).
68. Liu, X. et al. Engineering yeast for the production of breviscapine by genomic analysis and synthetic biology approaches. *Nat. Commun.* **9**, 448 (2018).
69. Zhao, Y. et al. Biosynthesis and biotechnological production of the anti-obesity agent celastrol. *Nat. Chem.* **15**, 1236–1246 (2023).
70. Zheng, H. et al. Transcriptional regulatory network of high-value active ingredients in medicinal plants. *Trends Plant Sci.* **28**, 429–446 (2023).
71. Liu, S., Zhang, Q., Kollie, L., Dong, J. & Liang, Z. Molecular networks of secondary metabolism accumulation in plants: current understanding and future challenges. *Ind. Crops Prod.* **201**, 116901 (2023).
72. Vom Endt, D., Soares e Silva, M., Kijne, J. W., Pasquali, G. & Memelink, J. Identification of a bipartite jasmonate-responsive promoter element in the Catharanthus roseus ORCA3 transcription factor gene that interacts specifically with AT-Hook DNA-binding proteins. *Plant Physiol.* **144**, 1680–1689 (2007).
73. Jiang, L. et al. Haplotype-resolved genome assembly of Bletilla striata (Thunb.) Reichb. f. to elucidate medicinal value. *Plant J.* **111**, 1340–1353 (2022).
74. Liu, S. et al. SmbHLH60 and SmMYC2 antagonistically regulate phenolic acids and anthocyanins biosynthesis in Salvia miltiorrhiza. *J. Adv. Res.* **42**, 205–219 (2022).
75. Jiang, T. et al. Transcription factor PgNAC72 activates DAMMARENEDIOL SYNTHASE expression to promote ginseng saponin biosynthesis. *Plant Physiol.* **195**, 2952–2969 (2024).
76. Shen, Q. et al. The jasmonate-responsive Aa MYC 2 transcription factor positively regulates artemisinin biosynthesis in Artemisia annua. *N. Phytol.* **210**, 1269–1281 (2016).
77. Zhang, H. et al. The basic helix-loop-helix transcription factor CrMYC2 controls the jasmonate-responsive expression of the ORCA genes that regulate alkaloid biosynthesis in Catharanthus roseus. *Plant J.* **67**, 61–71 (2011).
78. Du, T. et al. SmbHLH37 functions antagonistically with SmMYC2 in regulating jasmonate-mediated biosynthesis of phenolic acids in Salvia miltiorrhiza. *Front. Plant Sci.* **9**, 1720 (2018).
79. Sun, W. et al. Characterization of the horse chestnut genome reveals the evolution of aescin and aesculin biosynthesis. *Nat. Commun.* **14**, 6470 (2023).
80. Li, P. et al. Multiomics analyses of two Leonurus species illuminate leonurine biosynthesis and its evolution. *Mol. Plant* **17**, 158–177 (2024).
81. Li, C.-Y. et al. The sage genome provides insight into the evolutionary dynamics of diterpene biosynthesis gene cluster in plants. *Cell Rep.* **40**, 111236 (2022).
82. Rai, A. et al. Chromosome-scale genome assembly of Glycyrrhiza uralensis revealed metabolic gene cluster centred specialized metabolites biosynthesis. *DNA Res.* **29**, dsac043 (2022).
83. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* **45**, W55–W63 (2017).
84. Schläpfer, P. et al. Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiol.* **173**, 2041–2059 (2017).
85. Töpfer, N., Fuchs, L.-M. & Aharoni, A. The PhytoClust tool for metabolic gene clusters discovery in plant genomes. *Nucleic Acids Res.* **45**, 7049–7063 (2017).
86. Xiong, X. et al. The Taxu s genome provides insights into paclitaxel biosynthesis. *Nat. Plants* **7**, 1026–1036 (2021).
87. Kellner, F. et al. Genome-guided investigation of plant natural product biosynthesis. *Plant J.* **82**, 680–692 (2015).
88. Xu, H. et al. Analysis of the genome sequence of the medicinal plant Salvia miltiorrhiza. *Mol. Plant* **9**, 949–952 (2016).
89. Ma, Y. et al. Expansion within the CYP71D subfamily drives the heterocyclization of tanshinones synthesis in Salvia miltiorrhiza. *Nat. Commun.* **12**, 685 (2021).
90. Wink, M. Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. *Phytochemistry* **64**, 3–19 (2003).

91. Berman, P. et al. Parallel evolution of cannabinoid biosynthesis. *Nat. plants* **9**, 817–831 (2023).
92. Huang, X.-C. et al. The gradual establishment of complex coumarin biosynthetic pathway in Apiaceae. *Nat. Commun.* **15**, 6864 (2024).
93. Hu, J. et al. Functional divergence of CYP76AKs shapes the chemodiversity of abietane-type diterpenoids in genus Salvia. *Nat. Commun.* **14**, 4696 (2023).
94. Bryson, A. E. et al. Uncovering a miltiradiene biosynthetic gene cluster in the Lamiaceae reveals a dynamic evolutionary trajectory. *Nat. Commun.* **14**, 343 (2023).
95. Wang, Z. & Peters, R. J. Dynamic evolution of terpenoid biosynthesis in the Lamiaceae. *Mol. Plant* **16**, 963–965 (2023).
96. Liscombe, D. K., MacLeod, B. P., Loukanina, N., Nandi, O. I. & Facchini, P. J. Evidence for the monophyletic evolution of benzylisoquinoline alkaloid biosynthesis in angiosperms. *Phytochemistry* **66**, 1374–1393 (2005).
97. Lichman, B. R. et al. The evolutionary origins of the cat attractant nepetalactone in catnip. *Sci. Adv.* **6**, eaba0721 (2020).
98. Peng, Z. et al. Neofunctionalization of an OMT cluster dominates polymethoxyflavone biosynthesis associated with the domestication of citrus. *Proc. Natl. Acad. Sci. USA* **121**, e2321615121 (2024).
99. Wang, Y. et al. Deletion and tandem duplications of biosynthetic genes drive the diversity of triterpenoids in Aralia elata. *Nat. Commun.* **13**, 2224 (2022).
100. Shi, J., Tian, Z., Lai, J. & Huang, X. Plant pan-genomics and its applications. *Mol. Plant* **16**, 168–186 (2023).
101. Fang, S. et al. HERB: a high-throughput experiment-and reference-guided database of traditional Chinese medicine. *Nucleic Acids Res.* **49**, D1197–D1206 (2021).
102. Ren, G. et al. Large-scale whole-genome resequencing unravels the domestication history of Cannabis sativa. *Sci. Adv.* **7**, eabg2286 (2021).
103. Chen, J. et al. Whole-genome and genome-wide association studies improve key agricultural traits of safflower for industrial and medicinal use. *Hortic. Res.* **10**, uhad197 (2023).
104. Xiao, L., Tan, H. & Zhang, L. Artemisia annua glandular secretory trichomes: the biofactory of antimalarial agent artemisinin. *Sci. Bull.* **61**, 26–36 (2016).
105. Jiang, Z., Gao, W. & Huang, L. Tanshinones, critical pharmacological components in Salvia miltiorrhiza. *Front. Pharmacol.* **10**, 202 (2019).
106. Witherup, K. M. et al. Taxus spp. needles contain amounts of taxol comparable to the bark of Taxus brevifolia: analysis and isolation. *J. Nat. Prod.* **53**, 1249–1255 (1990).
107. Guo, L. et al. Effects of ecological factors on secondary metabolites and inorganic elements of Scutellaria baicalensis and analysis of geoherblism. *Sci. China Life Sci.* **56**, 1047–1056 (2013).
108. Shaw, R., Tian, X. & Xu, J. Single-cell transcriptome analysis in plants: advances and challenges. *Mol. Plant* **14**, 115–126 (2021).
109. Li, C. et al. Single-cell multi-omics in the medicinal plant Catharanthus roseus. *Nat. Chem. Biol.* **19**, 1031–1041 (2023).
110. Sun, S. et al. Single-cell RNA sequencing provides a high-resolution roadmap for understanding the multicellular compartmentation of specialized metabolism. *Nat. Plants* **9**, 179–190 (2023).

## Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2023YFA0914100).

## Author contributions

C.-Y.Y. supervised the study. L.C. collected data and literature and performed data analysis. L.C. and C.-Y.Y. wrote the manuscript. Z.W., Q.-H.Z., and M.Y. discussed the manuscript organization. Q.-H.Z. Edited the manuscript. All authors read and approved the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-57448-8.

**Correspondence** and requests for materials should be addressed to Chu-Yu Ye.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.