



Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research

1K Herb Genomes Project—Mini Review

The Last Decade of Medicinal Plant Genomics: Advances and Challenges

Junfeng Chen^{a, #}, Yuchen Zhang^{b, #}, Yajing Li^{a, #}, Yiru Liu^a, Qing Li^{a, c}, Zongyou Lv^a, Milen I. Georgiev^{d, e, *}, Pan Liao^{b, f, g, *}^a State Key Laboratory of Discovery and Utilization of Functional Components in Traditional Chinese Medicine, Institute of Chinese Materia Medica, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China^b Department of Biology, Hong Kong Baptist University, Hong Kong 999077, China^c Department of Pharmacy, the Second Affiliated Hospital of Naval Medical University, Shanghai 200003, China^d Laboratory of Metabolomics, The Stephan Angeloff Institute of Microbiology, Bulgarian Academy of Sciences, Plovdiv 4000, Bulgaria^e Center of Plant Systems Biology and Biotechnology, Plovdiv 4000, Bulgaria^f State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Hong Kong 999077, China^g Institute of Systems Medicine and Health Sciences, Hong Kong Baptist University, Hong Kong 999077, China

ARTICLE INFO

Article history:

Received 1 July 2025

Revised 1 September 2025

Accepted 25 September 2025

Keywords:

Medicinal plant

Genome sequencing

Active metabolites

Biosynthesis pathway

Multomics

ABSTRACT

Over the past decade (2014–2025), owing to increasing genome sequencing technologies and decreasing cost, research on medicinal plant genomes (MPGs) has expanded extensively. This advancement has significantly accelerated progress in multiple areas, including elucidating specialized metabolic pathways, conserving endangered species, advancing molecular breeding strategies, and accelerating metabolic engineering innovations. Here, we attempt to provide a contemporary view of MPGs, including analyses of sequenced species and assembly data, the advantages of illustrating the metabolic pathways of active compounds, and demonstrating how MPGs facilitate the metabolic biosynthesis characterization of medicinal plants.

© 2025 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

In the past decade, advancements in high-throughput sequencing technologies, such as long-read and high-accuracy platforms, combined with substantially reduced costs, have led to a surge in *de novo* plant genome sequencing. Consequently, the focus has broadened from primarily economically significant crops and model species to encompass a diverse array of plant species [1,2], with medicinal plants emerging as a key focus of interest. Since the papaya genome was sequenced in 2008, recognized as the first medicinal plant genome (MPG) [3], more than 400 plant genomes have been published. In contrast to major economic crops, medicinal plants exhibit extraordinary species diversity and widespread distribution across the plant kingdom, leading to significant genomic variations. This diversity poses substantial challenges for MPG

sequencing and assembly, including large genome sizes, high heterozygosity, elevated repeat content, and polyploidy [4]. Significant advancements and innovations in sequencing technologies, software, and increasingly powerful computing resources, including the integration of artificial intelligence (AI) [5], have effectively addressed these challenges, driving rapid progress in MPG research. For example, we can now address the challenges associated with giant genomes, such as those of medicinal plants in the genera *Paris* and *Fritillaria* within the Liliaceae family [6–8].

The primary goal in constructing genome maps of medicinal plants is to elucidate the biosynthetic pathways for active components in medicinal plants. Genome mining can provide critical insights into biosynthetic gene clusters (BGCs), tandem duplications, and syntenic regions, enabling the discovery of novel genes involved in the biosynthesis of specialized metabolites [9]. These genes primarily include genes encoding biosynthetic enzymes but can also encode transcription factors and related metabolite transporters [10–12]. Moreover, the integration of genome mining with taxonomic phylogeny, metabolic profiling, and comparative genomics provides profound insights into the genetic mechanisms

* Corresponding authors.

E-mail addresses: milengeorgiev@gbg.bg (M.I. Georgiev), panliao@hkbu.edu.hk (P. Liao).

These authors contributed equally to this work.

<https://doi.org/10.1016/j.eng.2025.09.024>

2095-8099/© 2025 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

underlying metabolic innovation and chemical diversity, driven by the interplay between the evolution of catalytic enzymes and genomic architecture.

MPGs have been reviewed extensively [13–17]; hence, in this review, we aim to provide a comprehensive overview of how sequenced MPGs have been utilized. We begin by reviewing the past decade of MPG sequencing, presenting a detailed overview of the current landscape of sequenced MPGs. We subsequently examine key perspectives through which genomic data can elucidate the biosynthetic pathways of bioactive compounds while addressing the challenges associated with deeper genomic applications. Finally, we present future perspectives in this field.

2. MPGs sequenced from 2014 to 2025

2.1. Progress in the genome sequencing of medicinal plants

As of August 2025, more than 1800 plant species had been sequenced [2]. According to these plant species, we defined 435 medicinal plant species in total according to the pharmacopoeias of several countries and included a variety of herbs used in ethnic and folk medicine traditions, covering regions such as Asia, Europe, the Americas, and Africa (Table S1 in Appendix A). From 2014 to 2019, only a limited number (1–12 species) of MPGs were published each year, primarily because of the high technical barriers and substantial sequencing costs associated with plant genome sequencing during this period. The subsequent adoption of advanced sequencing technologies, such as PacBio Sequel II's HiFi reads and Nanopore's ultralong reads, significantly accelerated the sequencing of plant genomes [18]. From 2022 to 2024, approximately 100 MPGs were published annually. Moreover, medicinal plants constitute an increasing proportion of newly sequenced plant species each year, reaching nearly 40% over the past three years (Fig. 1). With the rapid increase in the number of sequenced MPGs, assembly quality has significantly improved as well. The contig N50 size has increased sharply, from 1.66 Mb (2014–2020) to 13.17 Mb (2021–2023), and further to 47.05 Mb (2024–2025) on average. Although genome size and N50 size are not correlated, as sequencing technologies advance, it becomes easier to assemble large genomes. The largest MPG recently reported, also known as the largest sequenced plant genome, is that of *Paris polyphylla* (*P. polyphylla*), with a size of 54.58 Gb [7]. This milestone may pave the way for the sequencing of more giant MPGs in the near future, such as those of *Fritillaria* (exceeding 50 Gb in size) [5]. Notably, errors can occur during genome assembly and gene annotation. The application of third-generation transcriptomic techniques has been demonstrated to be useful for correcting and confirming gene sequences when combined with genome data. Another breakthrough enabled by sequencing technology is the use of telomere-to-telomere (T2T) gapless assemblies. Since 2023, a total of twelve T2T genomes of medicinal plants have been reported. However, owing to the doubled sequencing costs and extended analysis time, T2T has not yet been widely adopted in MPG sequencing.

In addition to their technical advantages, certain medicinal plant species of significant research interest have been repeatedly resequenced to improve assembly quality, meeting the demands of advanced research. Moreover, regular genome updates are highly necessary. For instance, at least six versions of the *Salvia miltiorrhiza* (*S. miltiorrhiza*) genome have been reported [19–22], with the latest representing a homozygous accession to support a highly efficient clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein 9 (Cas9)-mediated gene editing system [23]. However, uncertainty remains regarding the composition of the BGC responsible for tanshinone biosynthesis. Different genome versions reveal distinct gene loci, and we cannot

yet determine whether this is due to assembly errors or structural variations among samples (Table S2 in Appendix A). Similarly, four versions of the *Scutellaria baicalensis* (*S. baicalensis*) genome exist, with the most recent updated to a T2T assembly [24].

Medicinal plants are widely used worldwide. In addition to China, countries such as Japan, Korea, Vietnam, Singapore, Malaysia, Canada, Australia, the United Kingdom, Germany, France, the Netherlands, Switzerland, African countries, Russia, and Middle Eastern countries, as well as some states in the United States, also traditionally utilize medicinal plants. We further analyzed the countries and regions that completed these genome sequences and revealed that China contributed most of the MPG *de novo* assemblies (72.5%) in the past decade. There are several major explanatory reasons. First, the strong reliance on traditional Chinese medicine (TCM) drives demand for research into MPGs. Additionally, the economic value of medicinal plants has spurred related studies. Finally, the Chinese government's strong support for TCM policies has driven substantial investment in plant biology research associated with medicinal plants.

2.2. Database for MPGs

Specialized databases for MPGs serve as fundamental platforms for the systematic cataloguing and effective utilization of these hard-earned big data (Table 1, Table S3 in Appendix A). First, while not exclusively dedicated to MPGs, PubPlant compiles the most comprehensive information for published plant genome data and continues tracking and continuous updates, establishing it as the primary resource for accessing MPG information [2]. The most extensive high-throughput sequencing data, encompassing genome and transcriptome data, were collected for a total of 3257 medicinal plants. The platform links to the China National Gene-Bank Database (CNCBdb) for data download. The medicinal plant species collected on the website can be identified on the basis of their taxonomic order. Integrated Medicinal Plantomics (IMP) involves the collection of 84 high-quality genome assemblies and provides a one-stop mode for data access and analysis. Beyond downloading the genome files, users can perform expression profile analysis, employ the integrated genome viewer (IGV) browser, and access the transcriptional factors (TF) target finder online [16]. In addition to the specialized MPG database, the recently released the Plant Genomes for Comparative Phylogenomics (PGCG) consolidates the largest collection of 719 plant genomes, encompassing numerous medicinal plants [25]. All the genome annotations were uniformly reconstructed and standardized with PGCP-specific identifiers, enhancing data consistency and usability. This facilitates advanced complex analyses online, such as chromosome-level synteny comparisons across multiple genomes. We believe that this will be the future trend for the next-generation MPG database, enabling more convenient online analyses based on significantly accurate and comprehensive MPG data, such as phylogenetics, microsynteny, and prediction of gene clusters related to secondary metabolic pathways. A recently released database, the Gene-encoded Natural Diverse Components Repository (GNDC), which leverages the MPG database, provides a powerful platform for novel drug development and synthetic biology [26]. While it does not host genome data itself, it contains more than 234 million naturally diverse components encoded directly or indirectly by genes, serving as a critical resource for these fields.

3. Elucidation of secondary metabolic pathways on the basis of MPGs

The primary objective of MPG sequencing is to elucidate the biosynthetic pathways of bioactive metabolites. Before MPGs could

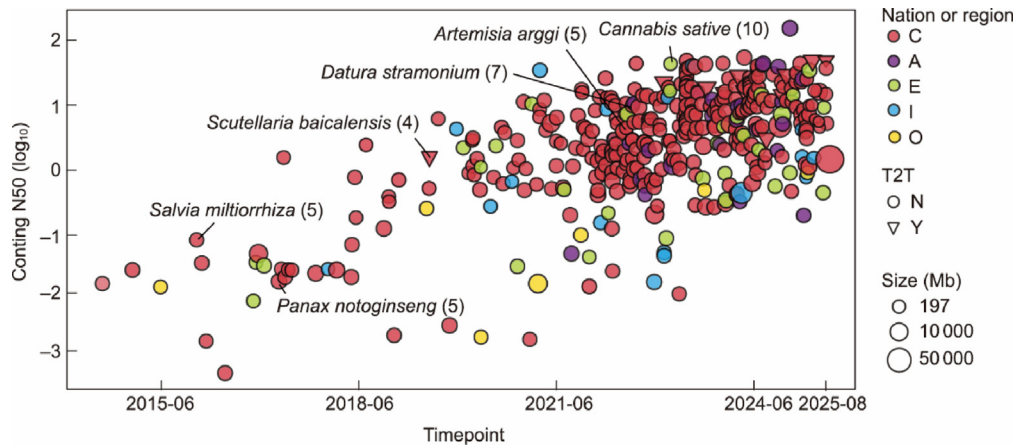


Fig. 1. Landscape of the MPG assembly. Assembly contiguity for 405 medicinal plant species with publicly available genome assemblies, spanning from 2014, is visualized on the basis of release date and contig N50 values ($\log_{10}$). Circles are colour-coded according to the nation or region responsible for generating the genome assembly and are scaled by genome size. Telomere-to-telomere (T2T) assemblies are indicated by triangles. Some selected species that have been sequenced multiple times are labelled with their species names, along with the number of sequencing instances indicated. The parameters included the date and contig N50 of the original version. The color codes are as follows: A, Americas; C, China; E, Europe; I, India; and O, other nations or regions.

Table 1

Database for the MPG.

Website	Number of collected genomes	Online functions	Country/region	Latest update
1 K Medicinal Plant Genome Database	113	a, b, c	China	2021
IBIS Databases	—	d	Australia	Continuing
Phytochemica	—	e	India	Continuing
Herbal Medicine Omics Database (MPOD)	50	a, b, c, h, i, k, l, m, n	China	2020
MPDB: Medicinal Plant DataBase (MPDB)	~200	a, b, f	China	Continuing
HerbMed	—	d	United States	2023
10KMP: Database of 10,000 Medicinal Plants	3113	a, f	China	2021
The Mint Genomics Resource (MGR)	6	a, b, g, l	United States	2021
IMP: Bridging the Gap for Medicinal Plant Genomics	84	a, b, c, f, g, i, j, l, m, n	China	2023
N3database	3517	a, b	China	2023
Kew Science	—	d	United Kingdom	Continuing
GNDC	—	e	China	Continuing

a: Genomic data, b: gene annotation (blast)/seq extraction, c: primer design, d: search species, e: phytochemicals, f: transcriptomes, g: pathways, h: biosynthetic tools, i: enrichment, j: differentially expressed genes analyses and mapping, k: phylogenetic tree, l: genome visualization, m: coexpression analysis, n: gene expression mapping.

be widely applied, conventional transcriptional coexpression appeared to be the most efficient approach for identifying candidate genes to elucidate metabolic pathways; however, identifying genes lacking transcriptional specificity or those within extensive paralogous gene families remains the key challenge [27]. Genome-wide analyses offer more comprehensive annotations, especially for tracing gene expansions, which serve as the primary genetic driving force of metabolic innovations in plants. A typical example is the analysis of the poppy genome, which revealed collinear segments on chromosomes 2 (Chr 2) and 11 (Chr 11). The segment on Chr 2 resulted from duplication, and a deletion event of the intergenic region between the P450 and oxidoreductase genes within this segment led to their fusion, resulting in the formation of a new gene, *STORR*. This enables poppy to specifically produce morphine [28,29] (Fig. 2(a)). Another advantage of genome-wide annotation is the ability to highlight BGCs. Genes located within or near these clusters are highly likely to be associated with the biosynthesis of specialized metabolites, encompassing not only catalytic enzyme-encoding genes but also potentially transcriptional regulatory factors and transport proteins. There are some well-known BGCs that are either conserved among plants or are taxon-specific, and hence facilitate the discovery of genes encoding specialized metabolites (Fig. 2(b)). For instance, the highly conserved diterpene BGC is widely distributed across the plant kingdom. In Poaceae and other cereals, this BGC mediates momilactones, specialized metabolites that inhibit competing

plant growth and microbial colonization through allelopathic effects [30]. In Lamiaceae, the core diterpene BGC, comprising a conserved terpene synthase (TPS), along with a *CYP76AH* and *CYP71D* gene array, likely originated in an early ancestor of the family (ca. 104 MYA) and is exceptionally well preserved in *Calli-carpa americana* (*C. americana*) [31,32]. This cluster enables the thorough elucidation of major diterpene scaffolds (kaurene, kola-venol, isokolavenol, and miltiradiene) [33], such as the diverse abietane diterpenoids across the genus *Salvia* [34–36] and clerodane diterpene biosynthesis in *Scutellaria barbata* (*S. barbata*) [30]. This BGC also drives specialized metabolites in ancient species, such as ginkgolide in *Ginkgo biloba* (*G. biloba*) [37]. A conserved BGC for triterpenoids, consisting of oxidosqualene cyclase (OSC) and multiple cytochrome P450 (CYP) genes, plays a pivotal role in the identification of specialized triterpenoid metabolites across diverse taxonomic groups. Notable examples include saponin biosynthesis in Sapindaceae [38,39], glycyrrhizin in *Glycyrrhiza uralensis* (*G. uralensis*) [40,41], and cucurbitacins in Cucurbitaceae [42]. Furthermore, a metabolic pathway may encompass multiple BGCs that form modules for distinct catalytic stages. Take the monoterpene indole alkaloid biosynthesis pathway as an instance. Three conserved BGCs have been discovered, including the module for iridoid biosynthesis [43], the BGC responsible for sesquiterpene skeleton oxidation formed by tandem duplication of CYPs [44], and the last module for condensation of tryptamine with secologanin to form strictosidine [45]. However, only a few of the

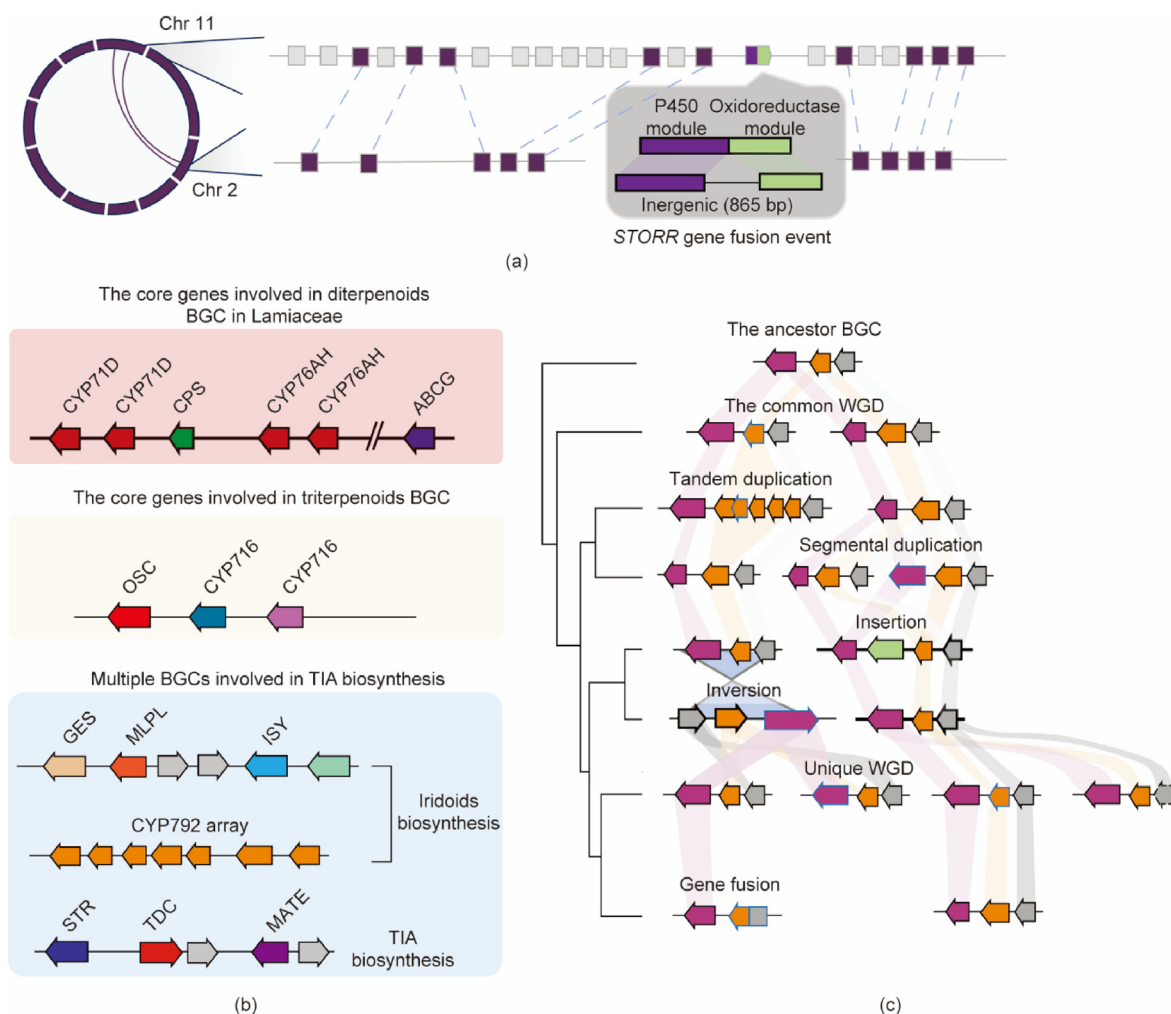


Fig. 2. Elucidation of secondary metabolic pathways based on MPGs. (a) Using the mechanism of the key morphine synthesis enzyme-encoding gene *STORR* in opium poppy as an example [28], this study demonstrates how genomic analysis can reveal novel catalytic enzyme-encoding genes. (b) Several important BGCs have been identified through MPG. (c) An evolutionary model summarizing the formation process of BGCs for specialized metabolites. CYP: cytochrome 450; CPS: copalyl diphosphate synthase; ABCG: ATP-binding cassette subfamily G transporter; OSC: oxidosqualene cyclase; GES: geraniol synthase; ISY: iridoid synthase; MLPL: major latex protein-like; STR: strictosidine synthase; TDC: tryptophan decarboxylase; MATE: multidrug and toxic compound extrusion; TIA: terpenoid indole alkaloid; WGD: whole-genome duplication.

currently identified BGCs are conserved across the plant lineage, with most being taxon-specific.

Although many gene clusters likely share a common ancestor, they have evolved significant lineage specificity to produce distinct specialized metabolites as species diverge. The growing abundance of MPGs offers a prime opportunity to elucidate the genetic mechanisms underlying specialized metabolites [9] (Fig. 2(c)). Integrating genome-wide gene annotation, extensive enzyme activity profiling, and phylogenetic analysis will further reveal how the evolution of BGCs drives specialized metabolism in plants, thus enabling the illustration of how genes within a conserved BGC evolve across diverse genomes, shaping specialized metabolite formation through subfunctionalization, neofunctionalization, or loss-of-function. Collective and comparative analysis enables the identification of structural variations in these loci within BGCs, such as gene insertions, segmental inversions, and tandem duplications, which act as pivotal driving forces of metabolic innovation. For instance, the BGC responsible for nepetalactone production has undergone a dynamic process of loss and re-emergence of iridoid biosynthesis in Nepetoideae, driven by the interplay between enzymes and genome evolution. The chronology of evolutionary events for nepetalactol-related short-chain reductase/dehydrogenases (*NEPS*) and progesterone 5 β -reductase (*PRISE*) loci, marked

by tandem duplication, positive selection, gene family expansion, and pseudogenization, collectively directed this prolonged and intricate origin, loss, and re-emergence of plant chemical diversity [43]. Given that specialized metabolites are typically taxonomically restricted, phylogenomic approaches are expected to offer robust evidence for elucidating the evolution of BGCs and their associated metabolic diversification. To demonstrate the pathway divergence of abietane-type diterpenoids across *Salvia* species, large-scale metabolic profiling combined with enzyme functional analysis, ancestral sequence reconstruction, and chemical trait mapping revealed that the loss of CYP76AK activity drives taxonomic chemical diversity among distinct *Salvia* lineages [36]. We anticipate that additional plant BGCs will be identified and that their distribution may be explored at other biological scales (e.g., single-cell or spatial levels) in future TCM research.

4. Multiomics approaches driven by MPGs

Leveraging genomic information as foundational data, advanced high-throughput sequencing techniques can be employed alongside genome mapping to dissect biosynthesis pathways [46]. Beyond elucidating these pathways, such approaches enable a deeper understanding of their regulatory mechanisms, shedding light

on the molecular basis of tissue- and organ-specific accumulation of active compounds, as well as the regulatory switches governing their expression across various developmental stages of the life cycle. Furthermore, these methods facilitate the investigation of variations in active component levels across different geographic regions and plant strains while also revealing the genetic mechanisms underlying responses to environmental changes.

4.1. Epigenomics

Beyond the transcriptional regulatory network, the epigenome likely manifests a pivotal upstream role in regulating metabolic biosynthesis. This regulation is typically elucidated through integrated analyses of histone methylation and acetylation, chromatin accessibility, and chromatin topology (Fig. 3(a)). However, such studies require extensive foundational research data, and their application in medicinal plants has not yet been widely adopted. For instance, in *Salvia* species, the BGC for abietane-type diterpenoids shows that genes for carnosic acid and tanshinones exhibit distinct transcriptional patterns in aerial and underground organs, which is consistent with the chromatin accessibility patterns revealed by the assay for transposase-accessible chromatin using sequencing (ATAC-seq) [47]. Comparative genomic and epigenomic analyses between diploid and tetraploid *Panax* species indicate that chromosomal interactions within chromosomes remain conserved after chromosome duplication. However, biased genetic fractionation and epigenetic regulation divergence during polyploidization processes drive remarkable biochemical diversity in specialized metabolites.

4.2. Population genetics

The availability of increasing MPGs has paved the way for population genetics studies, similar to those conducted for crops [48]. By identifying variations in bioactive compounds as quantitative traits, a meta-genome-wide association study (mGWAS) can identify critical genes that regulate the accumulation of these compounds, thereby elucidating the effects of germplasm, domestication, and environmental factors on medicinal plants (Fig. 3(b)). However, most medicinal plant species currently lack comprehensive germplasm collection, systematic breeding programs, and precise phenotypic assessments. This shortage of robust population samples represents a significant bottleneck for mGWAS in medicinal plants. Conversely, the extensive chemical and metabolic research available for nearly all medicinal plants provides a distinct advantage for conducting mGWAS. Several successful mGWAS based on MPGs serve as notable examples. For instance, mGWAS analysis of artemisinin content across different *Artemisia annua* populations revealed a positive correlation between the copy number of amorpho-4,11-diene synthase (*ADS*) gene loci and artemisinin levels, demonstrating a dosage effect of the *ADS* gene on artemisinin production [49]. Similarly, resequencing of *Perilla frutescens* populations revealed structural variations in the transcription factor MYB113, including long terminal repeat (LTR) insertions and exon fragment deletions, which altered the anthocyanin accumulation pattern and altered leaf color traits [50]. Additionally, genomic studies of related genera such as *Salvia*, *Papaver*, and *Panax* have laid the groundwork for pangenomic research on medicinal plants. However, such studies remain unreported, largely because of the need for more accurate and comprehensive population sample data to fully unlock the potential of these datasets. In addition to elucidating the biosynthesis of bioactive compounds, population genetics offers a powerful tool for investigating the agronomic traits of medicinal plants. For example, an integrated association analysis of environmental factors and medicinal quality across major *S. miltiorrhiza* production

regions in China revealed that a highly variable BGC (containing the tandemly arranged lignin synthase genes cinnamoyl-CoA reductase (*CCR*) and hydroxycinnamoyl-CoA shikimate/quinic acid hydroxycinnamoyl transferase (*HCT*), along with their transcription factor gene *WRKY40* significantly influenced root agronomic traits [51].

Moreover, in population genetics research, more powerful models, such as causal inference and Mendelian randomization, are becoming increasingly important methods for deciphering the genetic basis of complex traits, gene-environment interactions, and adaptive evolution mechanisms [52,53]. Although their application in plants remains in its early stages, the methodologies, informed by successful experiences in human genetics and epidemiology, hold promise for revealing their potential in studies on the evolution of medicinal plant germplasm and the impacts of environmental factors on quality variations.

4.3. Single-cell sequencing

Over the past decade, single-cell sequencing technology has advanced rapidly along with plant genome sequencing, enabling extraordinarily high-resolution analysis of plant secondary metabolic pathways [54,55]. Coupled with high-resolution metabolic characterization techniques, such as mass spectrometry imaging, this approach allows researchers to investigate the biosynthetic pathways of specialized metabolites at the tissue or cell level with unprecedented precision. Single-cell sequencing provides transcriptional profiles of distinct cell lineages, leveraging transcriptional heterogeneity to identify novel biosynthetic or regulatory genes. Fine-scale differentiation of cell lineages revealed variations in metabolic distribution (Fig. 3(c)). For example, the single-cell atlas of *Catharanthus roseus* leaves, the first published study for a medicinal plant, demonstrated that different modules of the terpenoid indole alkaloid (TIA) biosynthesis pathway exhibit unique transcriptional patterns across leaf cell lineages, indicating a compartmentalized model and suggesting the role of transporters in shuttling precursors between leaf tissues [56,57]. Furthermore, targeting specific biosynthetic cells enables the discovery of previously unknown genes. For instance, by isolating “hyper cells” from the single-cell atlas of *Hypericum perforatum* leaves, where hyperforin biosynthesis occurs, four prenyltransferases were precisely identified, leading to a complete understanding of the pathway [58]. Beyond elucidating the biosynthetic patterns of natural metabolites, researchers can perturb transcriptional variation within cell lineages to screen genes associated with metabolic pathways. A recent study on the paclitaxel pathway utilized this approach, combining multiplexed perturbation with single-nucleus RNA sequencing (mpXsn) to overcome challenges in discovering Taxol biosynthetic genes and identifying eight novel genes in the paclitaxel biosynthetic pathway [59]. In summary, single-cell sequencing is thus a powerful tool, providing precise insights for identifying synthase genes on the basis of whole-genome annotations. However, challenges remain, including inconsistent sequencing quality, difficulties in defining cell lineages in non-model species, and complex data analysis requirements.

5. Major challenges and perspectives

Breakthroughs in sequencing technologies have spurred an explosive increase in the *de novo* sequencing of MPGs. However, significant technical challenges persist for many species, primarily because of their complex genomic characteristics, such as high heterozygosity in wild populations, complex polyploidy, and the presence of exceptionally large genomes. Overcoming these hurdles requires continued improvements in sequencing technologies to achieve higher accuracy and enable longer read lengths, more efficient assembly methods and software, and reduced

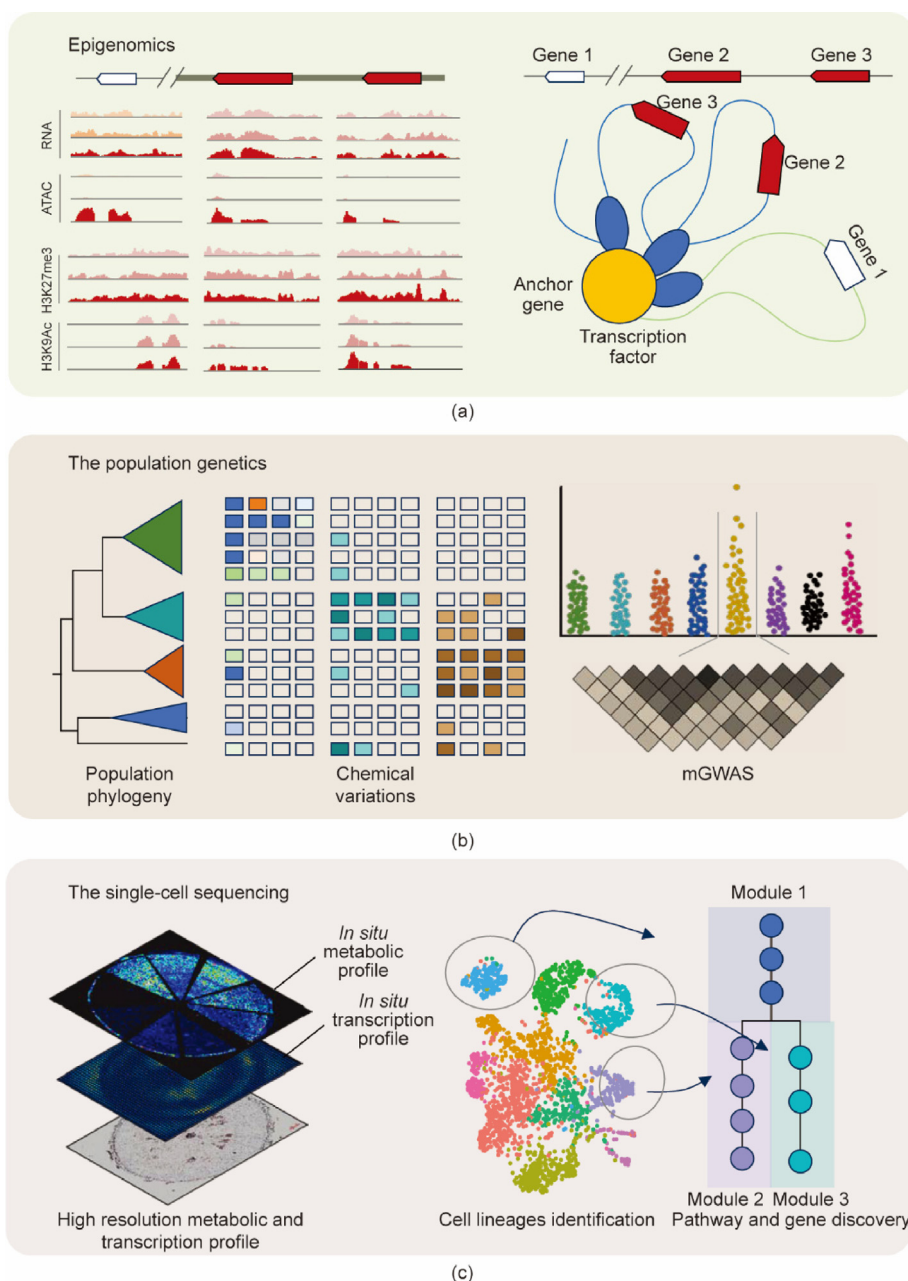


Fig. 3. How MPGs drive multiomics strategies to elucidate the biosynthesis of bioactive metabolites. (a) Integrating epigenomic analysis to elucidate the regulatory mechanisms of synthetic pathways. In the left panel, the representation reflects genome-based gene annotation. This allows us to conduct transcriptional analysis of catalytic enzyme-encoding genes, followed by investigations into their chromatin accessibility and histone methylation-related epigenomic regulatory mechanisms. The right panel illustrates the coregulation mechanisms of synthetic pathway genes through the study of chromatin spatial proximity. (b) Workflow for investigating the mechanisms underlying the differences in bioactive compounds within medicinal plant populations. (c) Integrating *in situ* metabolic characterization and single-cell sequencing technologies to characterize the cellular heterogeneity of bioactive compound biosynthetic pathways and to discover novel genes.

sequencing costs and time. In the near future, these advancements will likely enable the successful assembly of notoriously challenging MPGs, including highly sought-after medicinal plants in the Liliaceae family, such as *P. polyphylla* and *Fritillaria cirrhosa* (*F. cirrhosa*).

Although MPGs lead among all plant species in terms of *de novo* assembly, their practical applications lag behind those of major crops and model plant species. This is largely due to the vast diversity of medicinal plants, with only a few species benefiting from in-depth physiological and biochemical studies and a limited number of functionally characterized genes. Compared with major crops such as rice and tomato, there is a significant lack of well-

defined varieties and established genetic populations, which hinders the deeper utilization of MPGs in applications such as GWAS and pangenome analyses. Consequently, even with high-quality genomes, some critically important medicinal plants necessitate substantial future development, requiring extensive accumulation of germplasm and data to fully harness the potential of MPGs in elucidating physiological mechanisms and supporting breeding efforts.

While the acquisition of MPG has become increasingly streamlined, this has not simplified research on medicinal plants. In contrast, the availability of MPGs has enabled more complex and in-depth studies of metabolic pathways, significantly advancing the

application of multiomics in medicinal plant research. However, this requires researchers to possess strong genetic expertise and access to robust bioinformatics platforms, which can be daunting for beginners. To increase the accessibility of MPG utilization, several strategies can be adopted. First, improving the functionality of public databases by integrating more online analytical tools is essential, with some databases already making progress in this direction. Additionally, developing comprehensive and powerful analysis software, such as TBtools, which incorporates most genome analysis modules that do not require server computation and simplifies operations, enables researchers without coding skills to efficiently conduct analyses [60]. Looking ahead, the applications of AI and machine learning (ML) will undoubtedly play pivotal roles in automating genomic analysis. By developing large-scale genomic analysis models, AI can facilitate rapid analysis across every level and objective. Since 2020, AI has made significant advances in the field of plant genomics. Currently, the main challenges lie in three areas (Fig. 4): ① Tokenization of sequences. Different data formats can be used as inputs to deep learning models depending on the analytical objectives [61,62]. However, addressing specific requirements of medicinal MPGs, such as identifying regulatory sequences associated with organ-specific gene expression, demands the development of specialized models. ② Selection of deep learning models tailored to different analytical tasks. To date, transformer-based models have been the most widely adopted models in plant genomics research [63]. ③ AI applications in MPG span multiple levels. These include genome assembly, where current models enable automated processes from read splicing to chromosome scaffolding and correction [64,65]; sequence annotation, particularly in identifying repetitive elements and noncoding regions such as promoters; and enhancing

the efficiency of variant calling in population sequencing studies [66–68]. Ultimately, with advances in sequence interpretation and experimental validation, researchers will be able to design genomic sequences to precisely modulate gene expression patterns and engineer protein functions [67,69].

In summary, the rapid development of MPGs not only facilitates the physiological studies, breeding, and metabolic engineering of bioactive compounds in medicinal plants but also significantly broadens our understanding of the entire field of plant science.

CRediT authorship contribution statement

Junfeng Chen: Writing – original draft, Validation, Formal analysis, Data curation. **Yuchen Zhang:** Writing – original draft, Validation, Investigation. **Yajing Li:** Writing – review & editing, Formal analysis, Data curation. **Yiru Liu:** Investigation. **Qing Li:** Visualization. **Zongyou Lv:** Visualization. **Milen I. Georgiev:** Supervision, Project administration. **Pan Liao:** Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (82374159), the Chenguang Project of Shanghai (23CGA52), the Hong Kong Research Grants Council (RGC) Early Career Scheme 22100923, the NSFC/RGC Joint Research Scheme (N_HKBU201/23), Hong Kong Baptist University, Research Committee, Initiation Grant-Faculty Niche Research Areas (IG-FNRA) 2022/23 (RC-FNRA-IG/22-23/SCI/01), and the European Regional Development Fund through Programme Research Innovation and Digitalisation for Smart Transformation (BG16RFPR002-1.014-0003-C01).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.eng.2025.09.024>.

References

- [1] Marks RA, Hotelling S, Frandsen PB, VanBuren R. Representation and participation across 20 years of plant genome sequencing. *Nat Plants* 2021;7(12):1571–8.
- [2] Schwacke R, Bolger M, Usadel B. PubPlant—a continuously updated online resource for sequenced and published plant genomes. *Front Plant Sci* 2025;16:1603547.
- [3] Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 2008;452(7190):991–6.
- [4] Xie L, Gong X, Yang K, Huang Y, Zhang S, Shen L, et al. Technology-enabled great leap in deciphering plant genomes. *Nat Plants* 2024;10(4):551–66.
- [5] Zhang J, Yang Y, Si J, Chen D, Dong C, Chuan D, et al. Artificial intelligence in the discovery and modification of biological elements in medicinal plants. *Med Plant Biol* 2025;4(1):e012.
- [6] Li J, Lv M, Du L, Yunga A, Hao S, Zhang Y, et al. An enormous *Paris polyphylla* genome sheds light on genome size evolution and polyphyllin biogenesis. *bioRxiv* 2020:126920.
- [7] Xu S, Chen R, Zhang X, Wu Y, Yang L, Sun Z, et al. The evolutionary tale of lilies: giant genomes derived from transposon insertions and polyploidization. *The Innovation* 2024;5(6):100726.
- [8] Zeng P, Zong H, Han Y, Zhang W, Tian Z, Zhou B, et al. Two Melanthiaceae genomes with dramatic size difference provide insights into giant genome evolution and maintenance. *Nat Plants* 2025;11(8):1500–13.
- [9] Smit SJ, Lichman BR. Plant biosynthetic gene clusters in the context of metabolic evolution. *Nat Prod Rep* 2022;39(7):1465–82.

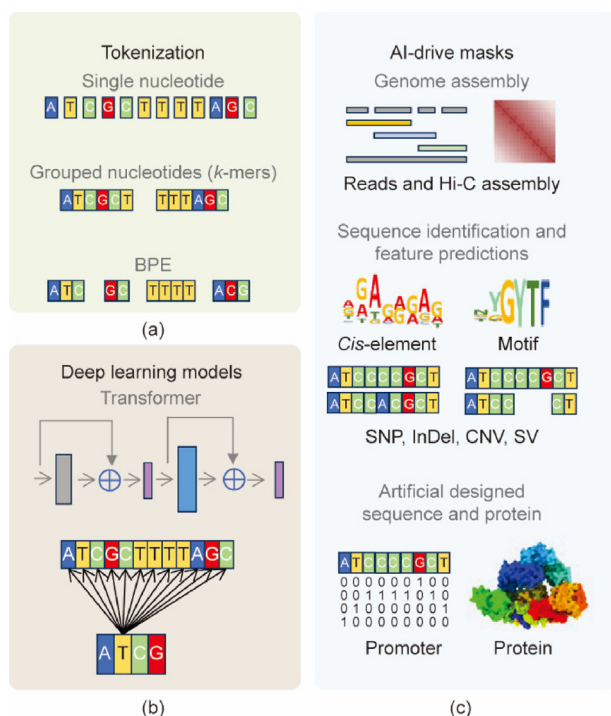


Fig. 4. Applications of AI in MPGs. (a) Genome sequence information can be tokenized at the single-nucleotide level, and multiple nucleotides can be grouped into *k*-mers or learned vocabularies (such as byte-pair encoding (BPE)). (b) Deep learning models operate on DNA sequence data with transformers as instances. (c) Analysis and in-depth mining of MPGs can be achieved with the applications of AI, including genome assembly, sequence identification, feature prediction, and AI-designed DNA and protein sequences. Hi-C: high-throughput/resolution chromosome conformation capture; SNP: single nucleotide polymorphism; InDel: insertion/deletion; CNV: copy number variant; SV: structural variant.

- [10] Shang Y, Ma Y, Zhou Y, Zhang H, Duan L, Chen H, et al. Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* 2014;346(6213):1084–8.
- [11] Zhong Y, Xun W, Wang X, Tian S, Zhang Y, Li D, et al. Root-secreted bitter triterpene modulates the rhizosphere microbiota to improve plant fitness. *Nat Plants* 2022;8(8):887–96.
- [12] Li Y, Chen J, Zhi J, Huang D, Zhang Y, Zhang L, et al. The ABC transporter SmABCG1 mediates tanshinones export from the peridermic cells of *Salvia miltiorrhiza* root. *J Integr Plant Biol* 2025;67(1):135–49.
- [13] Chakraborty P. Herbal genomics as tools for dissecting new metabolic pathways of unexplored medicinal plants and drug discovery. *Biochim Open* 2018;6:9–16.
- [14] Cheng QQ, Ouyang Y, Tang ZY, Lao CC, Zhang YY, Cheng CS, et al. Review on the development and applications of medicinal plant genomes. *Front Plant Sci* 2021;12:791219.
- [15] Su X, Yang L, Wang D, Shu Z, Yang Y, Chen S, et al. 1 K Medicinal Plant Genome Database: an integrated database combining genomes and metabolites of medicinal plants. *Hortic Res* 2022;9:uhac075.
- [16] Chen T, Yang M, Cui G, Tang J, Shen Y, Liu J, et al. IMP: bridging the gap for medicinal plant genomics. *Nucleic Acids Res* 2024;52(D1):D1347–54.
- [17] Cheng LT, Wang ZL, Zhu QH, Ye M, Ye CY. A long road ahead to reliable and complete medicinal plant genomes. *Nat Commun* 2025;16(1):2150.
- [18] Lang D, Zhang S, Ren P, Liang F, Sun Z, Meng G, et al. Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *GigaScience* 2020;9(12):giaa123.
- [19] Zhang G, Tian Y, Zhang J, Shu L, Yang S, Wang W, et al. Hybrid de novo genome assembly of the Chinese herbal plant danshen (*Salvia miltiorrhiza* Bunge). *GigaScience* 2015;4:62.
- [20] Xu H, Song J, Luo H, Zhang Y, Li Q, Zhu Y, et al. Analysis of the genome sequence of the medicinal plant *Salvia miltiorrhiza*. *Mol Plant* 2016;9(6):949–52.
- [21] Song Z, Lin C, Xing P, Fen Y, Jin H, Zhou C, et al. A high-quality reference genome sequence of *Salvia miltiorrhiza* provides insights into tanshinone synthesis in its red rhizomes. *Plant Genome* 2020;13(3):e20041.
- [22] Pan X, Chang Y, Li C, Qiu X, Cui X, Meng F, et al. Chromosome-level genome assembly of *Salvia miltiorrhiza* with orange roots uncovers the role of Sm20GD3 in catalyzing 15,16-dehydrogenation of tanshinones. *Hortic Res* 2023;10(6):uhad069.
- [23] Tian M, Luo L, Jin B, Liu J, Chen T, Tang J, et al. Highly efficient *Agrobacterium* rhizogenes-mediated gene editing system in *Salvia miltiorrhiza* inbred line bh2-7. *Plant Biotechnol J* 2025;23(6):2406–17.
- [24] Pei T, Zhu S, Liao W, Fang Y, Liu J, Kong Y, et al. Gap-free genome assembly and CYP450 gene family analysis reveal the biosynthesis of anthocyanins in *Scutellaria baicalensis*. *Hortic Res* 2023;10(12):uhad235.
- [25] Zhou X, Fan HY, Feng XY, Ruan Z, Yuan J, Han Q, et al. PGCP: a comprehensive database of plant genomes for comparative phylogenomics. *Plant Biotechnol J* 2025;23(7):2928–30.
- [26] Chen W, Yu Z, Leng L, Sun D, Liu H, Gong R, et al. Artificial intelligence-curated repository of gene-encoded natural diverse components from herbal medicines. *The Innovation* 2025:101011.
- [27] Swamidatta SH, Lichman BR. Beyond co-expression: pathway discovery for plant pharmaceuticals. *Curr Opin Biotechnol* 2024;88:103147.
- [28] Guo L, Winzer T, Yang X, Li Y, Ning Z, He Z, et al. The opium poppy genome and morphinan production. *Science* 2018;362(6412):343–7.
- [29] Zhang RG, Lu C, Li GY, Lv J, Wang L, Wang ZX, et al. Subgenome-aware analyses suggest a reticulate allopolyploidization origin in three *Papaver* genomes. *Nat Commun* 2023;14(1):2204.
- [30] Mao L, Kawaide H, Higuchi T, Chen M, Miyamoto K, Hirata Y, et al. Genomic evidence for convergent evolution of gene clusters for momilactone biosynthesis in land plants. *Proc Natl Acad Sci USA* 2020;117(22):12472–80.
- [31] Bryson AE, Lanier ER, Lau KH, Hamilton JP, Vaillancourt B, Mathieu D, et al. Uncovering a mitratriene biosynthetic gene cluster in the Lamiaceae reveals a dynamic evolutionary trajectory. *Nat Commun* 2023;14(1):343.
- [32] Wang Z, Peters RJ. Dynamic evolution of terpenoid biosynthesis in the Lamiaceae. *Mol Plant* 2023;16(6):963–5.
- [33] Li H, Wu S, Lin R, Xiao Y, Malaco Morotti AL, Wang Y, et al. The genomes of medicinal skullcaps reveal the polyphyletic origins of clerodane diterpene biosynthesis in the family Lamiaceae. *Mol Plant* 2023;16(3):549–70.
- [34] Scheler U, Brandt W, Porzel A, Rothe K, Manzano D, Božić D, et al. Elucidation of the biosynthesis of carnosic acid and its reconstitution in yeast. *Nat Commun* 2016;7(1):12942.
- [35] Ma Y, Cui G, Chen T, Ma X, Wang R, Jin B, et al. Expansion within the CYP71D subfamily drives the heterocyclization of tanshinones synthesis in *Salvia miltiorrhiza*. *Nat Commun* 2021;12(1):685.
- [36] Hu J, Qiu S, Wang F, Li Q, Xiang CL, Di P, et al. Functional divergence of CYP76AKs shapes the chemodiversity of abietane-type diterpenoids in genus *Salvia*. *Nat Commun* 2023;14(1):4696.
- [37] Forman V, Luo D, Geu-Flores F, Lemcke R, Nelson DR, Kampranis SC, et al. A gene cluster in *Ginkgo biloba* encodes unique multifunctional cytochrome P450s that initiate ginkgolide biosynthesis. *Nat Commun* 2022;13(1):5143.
- [38] Sun W, Yin Q, Wan H, Gao R, Xiong C, Xie C, et al. Characterization of the horse chestnut genome reveals the evolution of aescin and aesculin biosynthesis. *Nat Commun* 2023;14(1):6470.
- [39] Mai Y, Hu H, Ji W, Xiao Y, Zhou H, Zeng Z, et al. Evolution and functional characterization of a biosynthetic gene cluster for saponin biosynthesis in Sapindaceae. *Mol Plant* 2025;18(7):1089–93.
- [40] Mochida K, Sakurai T, Seki H, Yoshida T, Takahagi K, Sawai S, et al. Draft genome assembly and annotation of *Glycyrrhiza uralensis*, a medicinal legume. *Plant J* 2017;89(2):181–94.
- [41] Li Y, Xia C, Luo M, Huang Y, Xia Z, Li Y, et al. Comparative genomics of three medicinal *Glycyrrhiza* species unveiled novel candidates for the production of important bioactive compounds. *Plant J* 2025;122(4):e70223.
- [42] Zhou Y, Ma Y, Zeng J, Duan L, Xue X, Wang H, et al. Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. *Nat Plants* 2016;2(16183):16183.
- [43] Lichman BR, Godden GT, Hamilton JP, Palmer L, Kamileen MO, Zhao D, et al. The evolutionary origins of the cat attractant nepetalactone in catnip. *Sci Adv* 2020;6(20):eaba0721.
- [44] Rodríguez-López CE, Jiang Y, Kamileen MO, Lichman BR, Hong B, Vaillancourt B, et al. Phylogeny-aware chemoinformatic analysis of chemical diversity in Lamiaceae enables iridoid pathway assembly and discovery of aucubin synthase. *Mol Biol Evol* 2022;39(4):msac057.
- [45] Rai A, Hirakawa H, Nakabayashi R, Kikuchi S, Hayashi K, Rai M, et al. Chromosome-level genome assembly of *Ophiorrhiza pumila* reveals the evolution of camptothecin biosynthesis. *Nat Commun* 2021;12(1):405.
- [46] Sun W, Xu Z, Song C, Chen S. Herbgenomics: decipher molecular genetics of medicinal plants. *The Innovation* 2022;3(6):100322.
- [47] Li CY, Yang L, Liu Y, Xu ZG, Gao J, Huang YB, et al. The sage genome provides insight into the evolutionary dynamics of diterpene biosynthesis gene cluster in plants. *Cell Rep* 2022;40(7):111236.
- [48] Song B, Ning W, Wei D, Jiang M, Zhu K, Wang X, et al. Plant genome resequencing and population genomics: current status and future prospects. *Mol Plant* 2023;16(8):1252–68.
- [49] Liao B, Shen X, Xiang L, Guo S, Chen S, Meng Y, et al. Allele-aware chromosome-level genome assembly of *Artemisia annua* reveals the correlation between ADS expansion and artemisinin yield. *Mol Plant* 2022;15(8):1310–28.
- [50] Zhang Y, Shen Q, Leng L, Zhang D, Chen S, Shi Y, et al. Incipient diploidization of the medicinal plant *Perilla* within 10,000 years. *Nat Commun* 2021;12(1):5508.
- [51] Yu H, Liao J, Jiang Y, Zhong M, Tao S, Chai S, et al. Ecotype-specific phenolic acid accumulation and root softness in *Salvia miltiorrhiza* are driven by environmental and genetic factors. *Plant Biotechnol J* 2025;23(6):2224–41.
- [52] Xu F, Cheng Q, Liu S, Jiang S, Zhang J, Mao X, et al. MRBGR: a versatile toolbox for genetic regulation inference from population-scale multi-omics data. *Plant Commun* 2025;6(3):101197.
- [53] Kimmel K, Dee LE, Avolio ML, Ferraro PJ. Causal assumptions and causal inference in ecological experiments. *Trends Ecol Evol* 2021;36(12):1141–52.
- [54] Yu X, Liu Z, Sun X. Single-cell and spatial multi-omics in the plant sciences: technical advances, applications, and perspectives. *Plant Commun* 2023;4(3):100508.
- [55] Chen C, Zhang X, Yue M. Spatial multi-omics in medicinal plants: from biosynthesis pathways to industrial applications. *Trends Plant Sci* 2024;29(5):510–3.
- [56] Li C, Wood JC, Vu AH, Hamilton JP, Rodriguez Lopez CE, Payne RME, et al. Single-cell multi-omics in the medicinal plant *Catharanthus roseus*. *Nat Chem Biol* 2023;19(8):1031–41.
- [57] Sun S, Shen X, Li Y, Li Y, Wang S, Li R, et al. Single-cell RNA sequencing provides a high-resolution roadmap for understanding the multicellular compartmentation of specialized metabolism. *Nat Plants* 2023;9(1):179–90.
- [58] Wu S, Morotti ALM, Yang J, Wang E, Tatsis EC. Single-cell RNA sequencing facilitates the elucidation of the complete biosynthesis of the antidepressant hyperforin in *St. John's wort*. *Mol Plant* 2024;17(9):1439–57.
- [59] McClune CJ, Liu JC, Wick C, De La Peña R, Lange BM, Fordyce PM, et al. Discovery of *FoT01* and *Taxol* genes enables biosynthesis of baccatin III. *Nature* 2025;643(8071):582–92.
- [60] Chen C, Wu Y, Li J, Wang X, Zeng Z, Xu J, et al. TBtools-II: a “one for all, all for one” bioinformatics platform for biological big-data mining. *Mol Plant* 2023;16(11):1733–42.
- [61] Choi SR, Lee M. Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. *Biology* 2023;12(7):1033.
- [62] Consens ME, Dufault C, Wainberg M, Forster D, Karimzadeh M, Goodarzi H, et al. Transformers and genome language models. *Nat Mach Intell* 2025;7(3):346–62.
- [63] Farooq MA, Gao S, Hassan MA, Huang Z, Rasheed A, Hearne S, et al. Artificial intelligence in plant breeding. *Trends Genet* 2024;40(10):891–908.
- [64] Jiang Z, Peng Z, Wei Z, Sun J, Luo Y, Bie L, et al. A deep learning-based method enables the automatic and accurate assembly of chromosome-level genomes. *Nucleic Acids Res* 2024;52(19):e92.
- [65] Wen H, Yang J, Zhao X, Wang X, Lei J, Li Y, et al. TRFill: synergistic use of HiFi and Hi-C sequencing enables accurate assembly of tandem repeats for population-level analysis. *Genome Biol* 2025;26(1):227.
- [66] Avsec Ž, Latysheva N, Cheng J, Novati G, Taylor KR, Ward T. AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model. *bioRxiv* 2025.
- [67] Gosai SJ, Castro RI, Fuentes N, Butts JC, Moury K, Alasoadura M, et al. Machine-guided design of cell-type-targeting cis-regulatory elements. *Nature* 2024;634(8036):1211–20.
- [68] Wang D, Tan Z, Gao J, Zhang S, Shen J, Lu Y. AI4Protein: transforming the future of protein design. *Sci China Life Sci*. In press.
- [69] Fei H, Li Y, Liu Y, Wei J, Chen A, Gao C. Advancing protein evolution with inverse folding models integrating structural and evolutionary constraints. *Cell* 2025;188(17):4674–4692.e19.