Medicinal
Plant Biology

ARTICLE

Open Access

https://doi.org/10.48130/mpb-0024-0004
Medicinal Plant Biology **2024**, 3: e003

TCMPG 2.0: an enhanced database of traditional Chinese medicine plant genomes

Fanbo Meng[1,2], Tianzhe Chu[3], Lianjiang Hu[3], Mengqing Zhang[4], Qian Cheng[3], Xiuping Yang[3], Zhuo Liu[5], Yuannong Ye[6*], Xiaoming Song[5*] and Wei Chen[1,2*]

[1] *State Key Laboratory of Southwestern Chinese Medicine Resources, School of Basic Medicine, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China*
[2] *State Key Laboratory of Southwestern Chinese Medicine Resources, Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China*
[3] *School of Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China*
[4] *School of Intelligent Medicine, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China*
[5] *School of Life Sciences, North China University of Science and Technology, Tangshan 063210, China*
[6] *Bioinformatics and BioMedical Bigdata Mining Laboratory, School of Big Health, Guizhou Medical University, Guiyang 550025, China*
* Corresponding authors, E-mail: ynye.uestc@gmail.com; songxm@ncst.edu.cn; greatchen@ncst.edu.cn

Abstract

With the rapid advancements in sequencing technology, an increasing number of genome sequences for medicinal plants have become available. In 2021, TCMPG was introduced as a comprehensive database dedicated to traditional Chinese medicine (TCM) plant genomes, capturing significant attention within the scientific community. To offer invaluable resources to researchers, this database received an upgrade to its latest version, named TCMPG 2.0, accessible at http://cbcb.cdutcm.edu.cn/TCMPG2/. TCMPG 2.0 surpasses its predecessor by adding 114 medicinal plants, 129 genomes, and 159 related herbs. Recognizing the critical role of ingredients in assessing the pharmacological effects of Chinese medicines, TCMPG 2.0 has also amassed data on 13,868 herbal ingredients. Additionally, four practical analytical tools, including Heatmap, Primer3, PlantiSMASH, and CRISPRCasFinder, have been integrated into the toolbox. These valuable enhancements significantly enhance the database's utility for researchers in the field.

Citation: Meng F, Chu T, Hu L, Zhang M, Cheng Q, et al. 2024. TCMPG 2.0: an enhanced database of traditional Chinese medicine plant genomes. *Medicinal Plant Biology* 3: e003 https://doi.org/10.48130/mpb-0024-0004

Introduction

Over time, a shift has occurred in the storage of medicinal plant genomes and related information. Previously, these resources were predominantly scattered in fragmented databases, which posed obstacles to the systematic research and application of medicinal plants. Fortunately, this landscape has gradually evolved with the emergence of TCMPG[1] and other databases related to traditional Chinese medicine (TCM)[2,3]. By integrating diverse medicinal plant genome resources, such as NCBI and NGDC[4], TCMPG has established itself as the most comprehensive and accessible database in its respective domains. Its initial version encompasses data from 160 medicinal plants, 195 genomes, and 255 herbs. Additionally, TCMPG offered a suite of five commonly used bioinformatics analysis tools. These advancements have successfully attracted a substantial number of researchers from 92 countries and regions within the past year.

Advancements in sequencing technology, coupled with significant reductions in sequencing costs, have enabled the sequencing of a greater number of medicinal plant genomes, including *Artemisia annua*[5] and *Polygala tenuifolia*[6]. Simultaneously, the surging interest in traditional Chinese medicine has ignited widespread academic research into the therapeutic components of herbal ingredients. Prominent instances include ginsenoside Rh2, exhibiting promise in general tumor prevention[7], and costunolide, renowned for its excellent anti-inflammatory effects[8]. Researchers have diligently worked on the extraction and isolation of active compounds found in herbs, resulting in a wealth of data pertaining to the constituents associated with herbal medicines. However, this constituent data is often reported sporadically in the literature or dispersed across various TCM-related databases. The availability of comprehensive and accurate information regarding the constituents of Chinese medicines is of paramount importance for all facets of TCM research.

Hence, there arose a pressing need for a more comprehensive and accurate database housing relevant data and information. In TCMPG 2.0, the updated version, the original dataset has undergone experienced substantial expansion, along with the inclusion of a new data category known as 'ingredients'. Furthermore, four newly analytical tools, namely the Heatmap to draw heatmaps, Primer3 to design PCR primers, PlantiSMASH to view BGCs, and CRISPRCasFinder to search for CRISPR arrays and Cas proteins, have been incorporated. These enhancements have not only revitalized the latest version but have also elevated it a more valuable resource for researchers working on medicinal plants. The redesigned web interface can be accessed at http://cbcb.cdutcm.edu.cn/TCMPG2/.

Materials and methods

Data collection and processing

The genomic information of medicinal plants and their associated traditional Chinese medicines (TCMs) were collected using the same methodology as that employed in TCMPG[1]. In summary, we extracted information on reported medicinal species from plaBiPD (www.plaBiPD.de/index.ep) and filtered out those listed in TCMID[9,10]. The genomic data of the chosen medicinal plants was manually obtained from the links provided in the relevant publications. The ingredient information of TCMs was sourced from 11 established databases, including TCMSP[11], HERB2.0[12], ITCM[13], TCM2COVID[14], TCMIO[15], TCMSID[16], TCMID[9,10], TM-MC2.0[17], HIT2.0[18], TCM-Suite[19], and ETCM2.0[20]. To ensure data accuracy and eliminate redundancies, we consolidated information about Chinese medicines and their ingredients from these 11 databases. We consolidated duplicate pairs of herbs and ingredients obtained from various databases and documented the source names of these databases in TCMPG 2.0. Subsequently, we employed RDKit to handle the SMILES (Simplified Molecular Input Line Entry System) of the ingredients, and eliminated any illegitimate SMILES. Notably, we conducted a manual review of the data, addressing issues such as incorrect Chinese medicine names and inaccurately encoded ingredient names. Finally, we expanded the dataset by extracting additional information from PubChem using pubchempy. In this step, ingredient names were replaced with corresponding titles found in the PubChem title field. To streamline downstream analysis, we filtered the ingredients based on the collected herb names, retaining only those with isomeric SMILES notations. To enhance the user's comprehension of the chemical structure of the ingredients and facilitate differentiation, such as categorizing them by type and visualizing the arrangement of atoms and bonds within each compound, we employed SmilesDrawer (v1.0.10)[21] for visualizing all SMILES. This visual approach not only enhanced the understanding of the chemical makeup but also provided a valuable means for researchers to identify patterns, similarities, and differences among the compounds under investigation. Furthermore, ADMETlab2.0[22] was used to predict the absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of all the ingredients.

Configuration of tools

To provide a more user-friendly BLAST interface, we upgraded SequenceServer to version 2.0.0. Notably, we've restructured the database presentation, organizing it in a tree-like hierarchy based on species taxonomy. This organization significantly improves the accessibility and navigation of the database. In order to facilitate the generation of heatmaps, we introduced the Heatmap service using the PyComplexHeatmap[23] package in Python, which is a reliable tool for heatmap generation. To further empower users with Primer3 functionality[24,25], we've seamlessly integrated the primer3-py package into TCMPG2.0. This addition allows users to design primers for specific sequences, offering greater flexibility and control over their experiments. In the pursuit of identifying genome-wide biosynthetic gene clusters (BGCs), we've harnessed the capabilities of the PlantiSMASH[26] program. By using its default parameters, we can now effectively pinpoint all BGCs in chromosome-level genomes. This valuable feature aids researchers in the exploration of gene clusters responsible for the biosynthesis of various compounds. Moreover, to provide a comprehensive analysis, we've incorporated CRISPRCas-Finder[27] into TCMPG2.0. This tool enables the identification of CRISPR sequences and Cas proteins in user-submitted sequences, enhancing our capacity to analyze and annotate genomic data.

Results

Data updates and extensions

In TCMPG 2.0, we've significantly expanded and enriched our data resources, focusing on plant, genomic, and herbal data (Fig. 1). We've introduced 114 newly sequenced plants, spanning across 37 orders and 70 families, thereby broadening the diversity of species represented in the database. To provide users with comprehensive insights into each plant species, we've incorporated external links to iPlant, Plants of the World Online, and Encyclopedia Britannica on the plant details page. This enhances access to a wealth of information. Corresponding to the newly added plant species, we've gathered genomic data for 129 additional species, marking a substantial 66.2% increase in the number of genomes available. This genomic data includes 17 scaffold-level assemblies and 112 chromosome-level assemblies, with two being complete telomere-to-telomere (T2T) finished genomes. To address growing interest and research needs, we've introduced 159 new herbal entries into the database, increasing the total count of herbs to 414. In summary, TCMPG 2.0 represents a significant expansion of our dataset, demonstrating our commitment to providing a more comprehensive and valuable resource for researchers and users.

Addition of a new data field

We have incorporated a new data field, namely the ingredient information of herbs, into TCMPG 2.0. To compile this valuable resource, we primarily gathered ingredient data from 11 existing databases (see Materials and methods). Following meticulous data integration and screening procedures, we successfully incorporated 13,868 unique ingredients associated with 397 herbs. This process yielded a total of 31,517

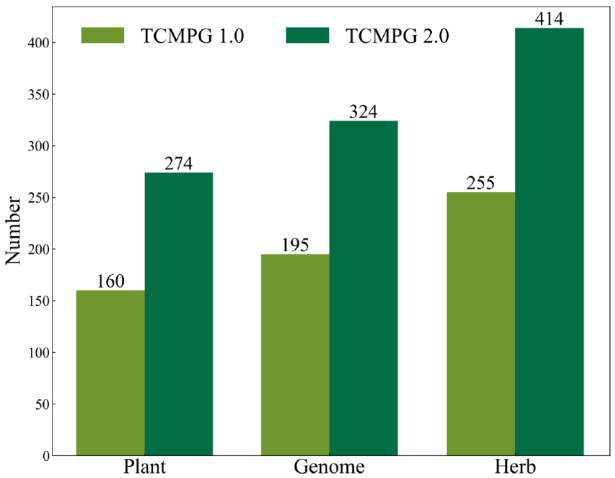

Fig. 1 Comparative results for the quantities of plants, genomes and herbs in TCMPG 1.0 and TCMPG 2.0.

records detailing the relationships between herbs and their respective ingredients (Fig. 2). Additionally, we harnessed the capabilities of ADMETlab 2.0 to predict the medicinal chemistry and ADMET properties of all the ingredients. Furthermore, we included records of associated ingredients on the herb page and records of related herbs on the ingredient page. These enhancements contribute to a more extensive and valuable information for users.

Enhancements and new additions to tools

Within TCMPG 2.0, we've made substantial upgrades and additions to our suite of tools to better cater to research needs. Notable improvements include the enhanced BLAST service, which has been fine-tuned to accommodate the ever-growing array of databases. We've gone a step further by reorganizing the existing database into a species taxonomy-based tree structure, making data navigation more intuitive and efficient. Additionally, to enrich the exploration of gene functions, we've introduced the sequence of *Arabidopsis thaliana* into our database, despite it not being a medicinal plant. Additionally, we have added simple sequence repeats (SSRs) identified from whole genome sequences to the SSR Finder.

In order to provide users with more practical tools for bioinformatics analysis, we have added four new tools to TCMPG 2.0 (Fig. 3). We offer a service called Heatmap, designed to assist users in generating heatmaps for expression profile data (Fig. 3a). Heatmap allows for the quick creation of heatmaps by simply uploading a data matrix. Additionally, users have the flexibility to customize their heatmaps by adjusting options such as color schemes and clustering methods based on their preferences. To aid experimentalists in PCR primer design, we have incorporated the essential features of Primer3 (Fig. 3b). Users can conveniently input the sequence into the designated box to swiftly generate primers. We have also conducted an analysis of secondary metabolic gene clusters (BGCs) across all chromosome-level genomes (Fig. 3c). Leveraging the capabilities of the PlantiSMASH tool, we've cataloged all BGCs, offering users a convenient method to explore gene clusters of interest. Moreover, we embedded CRISPRCasFinder to identify CRISPR arrays and Cas proteins in sequences (Fig. 3d). These enhancements collectively contribute to a more robust and user-friendly suite of tools, aimed at empowering researchers and experimentalists in their scientific endeavors.

Discussion

TCMPG has demonstrated its increasing significance as a bridge connecting traditional Chinese medicine with modern research. In the latest version of TCMPG, we've seen remarkable growth, with an expanded collection encompassing 274 plants, 324 genomes, and 414 herbs. We have incorporated information on 13,686 constituents found in 397 herbs. It is worth noting that certain ingredients lacking precise SMILES representation have not been included in TCMPG 2.0. Furthermore, we have integrated four new bioinformatics tools into TCMPG 2.0. These tools are not only commonly used but also powerful, making them highly valuable for experimental biologists. With the continuous release of more genomes, particularly the T2T version, we will persist in gathering new genomic data on medicinal plants and storing them in TCMPG. In addition to herbs, TCM contains animals and fungi, and many of their genomic data have already been published, such as, *Bungarus multicinctus*[28] and *Ganoderma lucidum*[29,30]. To expand the applicability of our database, future updates will incorporate the genomes of medicinal animals and fungi. Simultaneously, in view of the increasing number of studies focusing on TCM genomes and their analytical tools[31], we're committed to incorporating additional valuable tools into the TCMPG platform, ensuring that it remains a dynamic and indispensable resource for researchers in the field of traditional Chinese medicine and beyond.

Conclusions

TCMPG 2.0 stands as a dedicated effort towards building a more comprehensive and integrated database of medical plants. Our primary objective is to accelerate the modernization and standardization of traditional Chinese medicine. The recent enhancements and updates made to TCMPG have substantially improved its functionality while carefully maintaining the integrity of the original network structure. We firmly believe that TCMPG 2.0 will prove invaluable to the scientific

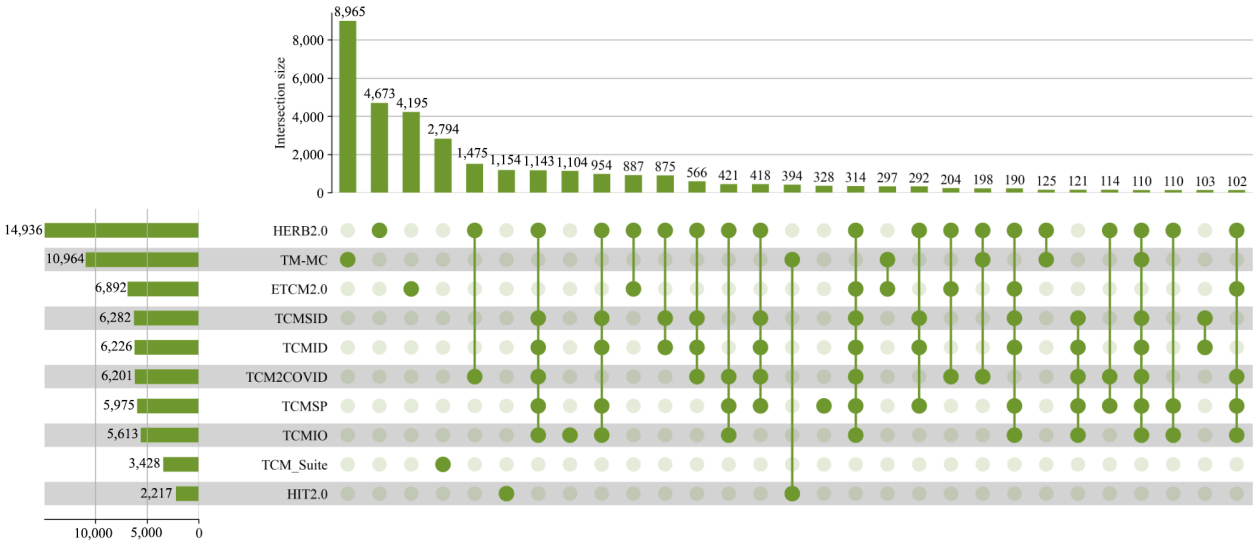

Fig. 2 Source databases for herb-ingredients in TCMPG 2.0 with more than 100 items.

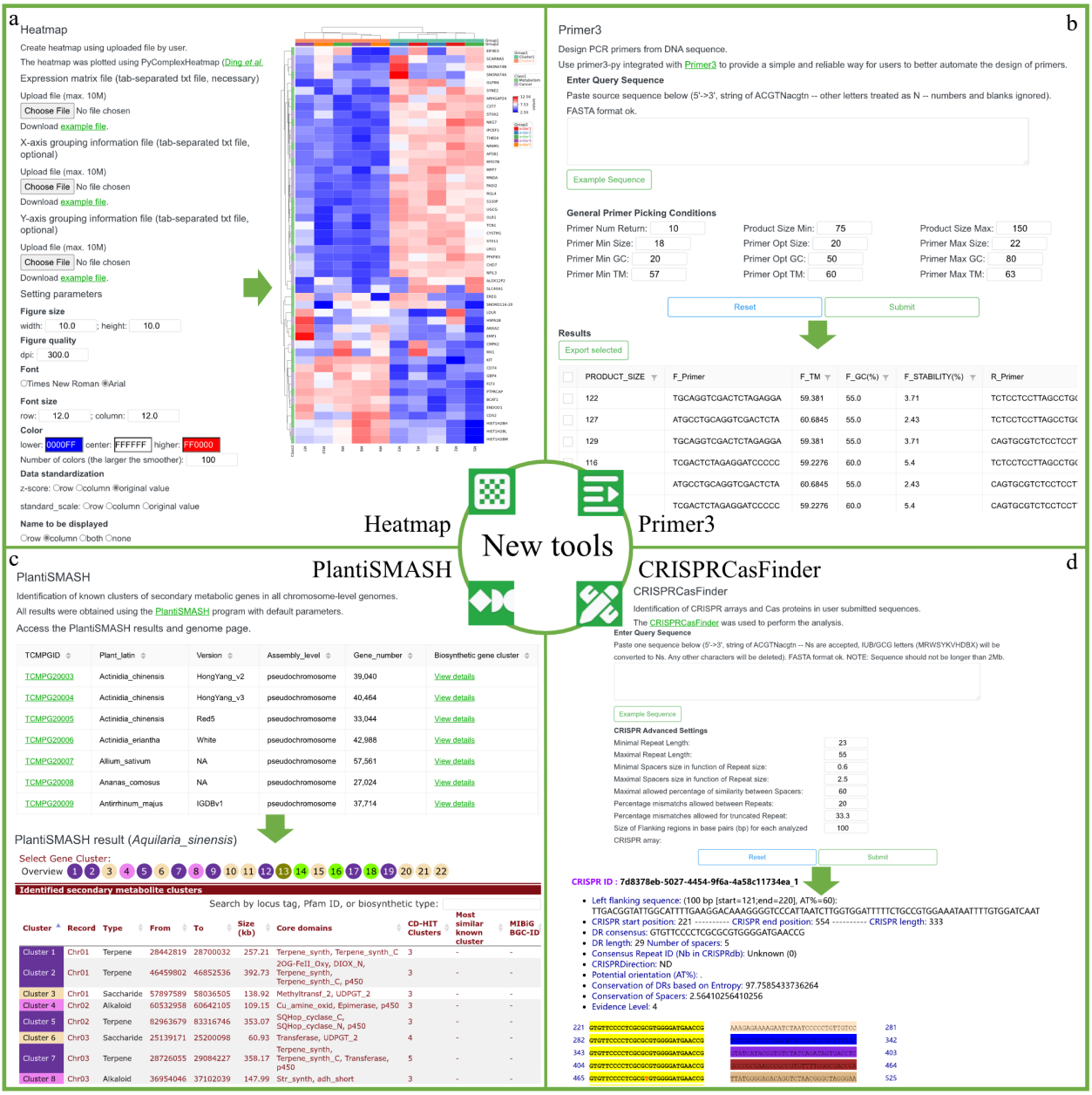

Fig. 3 The newly added tools in TCMPG 2.0.

community. It not only offers a wealth of information but also provides enhanced services, fostering a more in-depth exploration of traditional Chinese medicine. This platform is poised to play a pivotal role in advancing our understanding of traditional Chinese medicine in the context of modern research.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: Chen W, Song X, Ye Y; data collection and analysis: Meng F, Chu T, Hu L, Zhang M, Cheng Q, Yang X, Liu Z; analysis and interpretation of results: Meng F, Chu T, Hu L, Liu Z; draft manuscript preparation: Meng F, Chen W, Song X. All authors reviewed the results and approved the final version of the manuscript.

Data availability

All the data associated with this study are provided at http://cbcb.cdutcm.edu.cn/TCMPG2/.

Acknowledgments

This work was supported by Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (No: ZYYCXTD-D-202209), and the 'Xinglin Scholar' Discipline Talent Research Promotion Program of Chengdu University of TCM (No. MPRC2021036).

Conflict of interest

The authors declare that they have no conflict of interest.

Dates

Received 22 October 2023; Accepted 19 January 2024; Published online 6 February 2024

References

1. Meng F, Tang Q, Chu T, Li X, Lin Y, et al. 2022. TCMPG: an integrative database for traditional Chinese medicine plant genomes. *Horticulture Research* 9:uhac060
2. He S, Yang L, Ye S, Lin Y, Li X, et al. 2022. MPOD: Applications of integrated multi-omics database for medicinal plants. *Plant Biotechnology Journal* 20:797–99
3. Su X, Yang L, Wang D, Shu Z, Yang Y, et al. 2022. 1 K Medicinal Plant Genome Database: an integrated database combining genomes and metabolites of medicinal plants. *Horticulture Research* 9:uhac075
4. Members C-N, Partners. 2021. Database resources of the national genomics data center, china national center for bioinformation in 2021. *Nucleic Acids Research* 49:D18–D28
5. Liao B, Shen X, Xiang L, Guo S, Chen S, et al. 2022. Allele-aware chromosome-level genome assembly of *Artemisia annua* reveals the correlation between *ADS* expansion and artemisinin yield. *Molecular Plant* 15:1310–28
6. Meng F, Chu T, Feng P, Li N, Song C, et al. 2023. Genome assembly of *Polygala tenuifolia* provides insights into its karyotype evolution and triterpenoid saponin biosynthesis. *Horticulture Research* 10:uhad139
7. Wang YS, Chen C, Zhang SY, Li Y, Jin YH. 2021. (20S) Ginsenoside Rh2 Inhibits STAT3/VEGF Signaling by Targeting Annexin A2. *International Journal of Molecular Sciences* 22:9289
8. Xu H, Chen J, Chen P, Li W, Shao J, et al. 2023. Costunolide covalently targets NACHT domain of NLRP3 to inhibit inflammasome activation and alleviate NLRP3-driven inflammatory diseases. *Acta Pharmaceutica Sinica B* 13:678–93
9. Xue R, Fang Z, Zhang M, Yi Z, Wen C, et al. 2013. TCMID: Traditional Chinese Medicine integrative database for herb molecular mechanism analysis. *Nucleic Acids Research* 41:D1089–D1095
10. Huang L, Xie D, Yu Y, Liu H, Shi Y, et al. 2018. TCMID 2.0: a comprehensive resource for TCM. *Nucleic Acids Research* 46:D1117–D1120
11. Ru J, Li P, Wang J, Zhou W, Li B, et al. 2014. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. *Journal of Cheminformatics* 6:13
12. Fang S, Dong L, Liu L, Guo J, Zhao L, et al. 2021. HERB: a high-throughput experiment- and reference-guided database of traditional Chinese medicine. *Nucleic Acids Research* 49:D1197–D1206
13. Tian S, Zhang J, Yuan S, Wang Q, Lv C, et al. 2023. Exploring pharmacological active ingredients of traditional Chinese medicine by pharmacotranscriptomic map in ITCM. *Briefings in Bioinformatics* 24:bbad027
14. Ren L, Xu Y, Ning L, Pan X, Li Y, et al. 2022. TCM2COVID: A resource of anti-COVID-19 traditional Chinese medicine with effects and mechanisms. *iMeta* 1:e42
15. Liu Z, Cai C, Du J, Liu B, Cui L, et al. 2020. TCMIO: A comprehensive database of traditional Chinese medicine on immuno-oncology. *Frontiers in Pharmacology* 11:439
16. Zhang LX, Dong J, Wei H, Shi SH, Lu AP, et al. 2022. TCMSID: a simplified integrated database for drug discovery from traditional chinese medicine. *Journal of Cheminformatics* 14:89
17. Kim SK, Nam S, Jang H, Kim A, Lee JJ. 2015. TM-MC: a database of medicinal materials and chemical compounds in Northeast Asian traditional medicine. *BMC Complementary and Alternative Medicine* 15:218
18. Yan D, Zheng G, Wang C, Chen Z, Mao T, et al. 2022. HIT 2.0: an enhanced platform for Herbal Ingredients' Targets. *Nucleic Acids Research* 50:D1238–D1243
19. Yang P, Lang J, Li H, Lu J, Lin H, et al. 2022. TCM-Suite: A comprehensive and holistic platform for Traditional Chinese Medicine component identification and network pharmacology analysis. *iMeta* 1:e47
20. Zhang Y, Li X, Shi Y, Chen T, Xu Z, et al. 2023. ETCM v2.0: An update with comprehensive resource and rich annotations for traditional Chinese medicine. *Acta Pharmaceutica Sinica B* 13:2559–71
21. Probst D, Reymond JL. 2018. SmilesDrawer: Parsing and drawing SMILES-encoded molecular structures using client-side JavaScript. *Journal of Chemical Information and Modeling* 58:1–7
22. Xiong G, Wu Z, Yi J, Fu L, Yang Z, et al. 2021. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Research* 49:W5–W14
23. Ding W, Goldberg D, Zhou W. 2023. PyComplexHeatmap: A Python package to visualize multimodal genomics data. *iMeta* 2:e115
24. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, et al. 2012. Primer3-new capabilities and interfaces. *Nucleic Acids Research* 40:e115
25. Koressaar T, Remm M. 2007. Enhancements and modifications of primer design program Primer3. *Bioinformatics* 23:1289–91
26. Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH. 2017. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Research* 45:W55–W63
27. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, et al. 2018. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research* 46:W246–W251
28. Xu J, Guo S, Yin X, Li M, Su H, et al. 2023. Genomic, transcriptomic, and epigenomic analysis of a medicinal snake, *Bungarus multicinctus*, to provides insights into the origin of Elapidae neurotoxins. *Acta Pharmaceutica Sinica B* 13:2234–49
29. Chen S, Xu J, Liu C, Zhu Y, Nelson DR, et al. 2012. Genome sequence of the model medicinal mushroom *Ganoderma lucidum*. *Nature Communications* 3:913
30. Sonets IV, Dovidchenko NV, Ulianov SV, Yarina MS, Koshechkin SI, et al. 2023. Unraveling the polysaccharide biosynthesis potential of *Ganoderma lucidum*: A chromosome-level assembly using Hi-C sequencing. *Journal of Fungi* 9:1020
31. Chen Z, Li J, Hou N, Zhang Y, Qiao Y. 2021. TCM-Blast for traditional Chinese medicine genome alignment with integrated resources. *BMC Plant Biology* 21:339