SCIENCE CHINA
Life Sciences

•RESEARCH PAPER•

April 2022 Vol.65 No.4: 809–817
https://doi.org/10.1007/s11427-021-1968-7

Global Pharmacopoeia Genome Database is an integrated and mineable genomic database for traditional medicines derived from eight international pharmacopoeias

Baosheng Liao[1†], Haoyu Hu[1†], Shuiming Xiao[1†], Guanru Zhou[2], Wei Sun[1], Yang Chu[1], Xiangxiao Meng[1], Jianhe Wei[3], Han Zhang[4,5], Jiang Xu[1*] & Shilin Chen[1*]

[1]*Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China;*
[2]*Institute of Pharmacy, Hubei University of Chinese Medicine, Wuhan 430065, China;*
[3]*Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Beijing 100193, China;*
[4]*Pharmacy College, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China;*
[5]*Institute of Traditional Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin 301617, China*

Received April 8, 2021; accepted June 22, 2021; published online August 6, 2021

Genomic data have demonstrated considerable traction in accelerating contemporary studies in traditional medicine. However, the lack of a uniform format and dispersed storage limits the full potential of herb genomic data. In this study, we developed a Global Pharmacopoeia Genome Database (GPGD). The database contains 34,346 records for 903 herb species from eight global pharmacopoeias (Brazilian, Egyptian, European, Indian, Japanese, Korean, the Pharmacopoeia of the People's Republic of China, and U.S. Pharmacopoeia's Herbal Medicines Compendium). In particular, the GPGD contains 21,872 DNA barcodes from 867 species, 2,203 organelle genomes from 674 species, 55 whole genomes from 49 species, 534 genomic sequencing datasets from 366 species, and 9,682 transcriptome datasets from 350 species. Among the organelle genomes, 534 genomes from 366 species were newly generated in this study. Whole genomes, organelle genomes, genomic fragments, transcriptomes, and DNA barcodes were uniformly formatted and arranged by species. The GPGD is publicly accessible at http://www.gpgenome.com and serves as an essential resource for species identification, decomposition of biosynthetic pathways, and molecular-assisted breeding analysis. Thus, the database is an invaluable resource for future studies on herbal medicine safety, drug discovery, and the protection and rational use of herbal resources.

Global Pharmacopoeia Genome Database, herb, traditional medicine, genomics

Citation: Liao, B., Hu, H., Xiao, S., Zhou, G., Sun, W., Chu, Y., Meng, X., Wei, J., Zhang, H., Xu, J., et al. (2022). Global Pharmacopoeia Genome Database is an integrated and mineable genomic database for traditional medicines derived from eight international pharmacopoeias. Sci China Life Sci 65, 809–817. https://doi.org/10.1007/s11427-021-1968-7

INTRODUCTION

Traditional medicines have been globally used to treat various diseases for thousands of years, and critically, still play important roles in modern-day healthcare. They also serve as primary medicinal resources for approximately 85% of the world's population (Pešić, 2015) and also drug discovery, as approximately 46% of newly approved drugs between 1981 and 2019 were derived from natural sources or natural product mimics (Newman and Cragg, 2020). During the coronavirus disease 2019 (COVID-19) global pandemic, traditional medicines displayed encouraging effects in improving symptom management, reducing deterioration,

†Contributed equally to this work
*Corresponding authors (Jiang Xu, email: jxu@icmm.ac.cn; Shilin Chen, email: slchen@icmm.ac.cn)

mortality, and disease recurrence rates in China (Luo et al., 2020). Currently, more than 60 official pharmacopoeias exist in different countries and regions (World Health Organization, 2019), with 42 (70%) containing herb or natural product records.

However, toxic compounds (Chen et al., 2014; Gilbert, 2011) and counterfeits in the market (Chan, 2002; Han et al., 2016; Poon et al., 2006; Lowe et al., 2005) have increased concerns regarding the quality and safety of traditional medicines. Thus, the accurate identification of medicinal species, as a prerequisite for medication safety, is crucial for production and market regulation. DNA barcodes, which are short highly variable DNA fragments among species (Hebert et al., 2003), provide a practical solution to facilitate species authentication (Chen et al., 2010). A DNA barcode database (ITS2 and *psb*A-*trn*H, short DNA sequences from nucleus or chloroplast genomes with high efficiency for species identification) for herbal materials (http://www.tcmbarcode.cn/) was previously established (Chen et al., 2014; Chen et al., 2010) and is widely used in traditional medicine industries and research. In addition, combining the optimal use of single-locus barcodes and chloroplast genomes (super-barcodes) could be a new approach for efficient plant identification (Li et al., 2015). The whole plastid genome or its selected loci have been tested or validated in a range of herbs and plants (Chen et al., 2018; Guo et al., 2019; Shen et al., 2017b). The rich resources from herb species and the increasing need for herb identification require the simultaneous development of efficient and sufficient DNA barcodes/super-barcodes. Thus, a database integrating DNA barcodes and organelle genomes is an ideal solution.

Unlike synthetic drugs, medicinal plants are composed of complex natural compounds, which makes them crucial resources for drug discovery. Ingredients with significant medicinal activities have been extracted from herbs, e.g., artemisinin (Callaway and Cyranoski, 2015; Coordinating Research Group for the Structure of Artemisinin, 1977), paclitaxel (Weaver, 2014), and vinblastine (Muniraj et al., 2019). With the assistance of "omics" data, the decomposition of medicinal compound biosynthetic pathways has reached an unprecedented level (Guo et al., 2018; Xu et al., 2017). Molecular information on pharmacodynamic components from transcriptomic and genomic plant data has contributed to the fermentation engineering of desired chemical compounds (Atanasov et al., 2015; Smanski et al., 2016). To date, more than 60 herbal genomes have been sequenced (Table S1 in Supporting Information). Herb genomic data are growing rapidly; however, this growth is accompanied by several issues exemplified by the hosting of data by different research groups, various and non-uniform data formats, and unstable web services, which combined, create obstacles for data use. Additionally, ever-increasing versions of genome assemblies for a single species are constantly being generated thanks to continued improvements to previous genome assemblies and new cultivar sequencing strategies (Kim et al., 2018; Xu et al., 2017). Multiple versions of a genome assembly are likely to cause confusion for genomic data use. Therefore, the prevailing disordered state of genomic data use must be comprehensively processed and structured for future use.

Superior germplasm resources are crucial for herbal material production. However, most herbal products still rely on wild resources (Xiao et al., 2009), and many species have become endangered due to habitat destruction and extensive exploitation (Bantawa et al., 2009; Chik et al., 2015; Guan et al., 2002; Liang et al., 2008; Lu and Lan, 2013). The cultivation of wild medicinal herbs has become an imperative trend for sustainable development (Chen et al., 2005; Cordell, 2015). Traditional selection for plant species is uneconomical and time-consuming while molecular-assisted breeding based on genomic data considerably reduces the amount of breeding work (Geuna, 2017), which rapidly enriches germplasm resources and protects wild resources, as evidenced for *Panax notoginseng* (Dong et al., 2017) and *Perilla frutescens* (Shen et al., 2017a). Comprehensive "omics" databases for crop breeding programs have been established for rice (Tareke Woldegiorgis et al., 2019), maize (Portwood et al., 2019), and wheat (Alaux et al., 2018) and have been used globally for several years. However, no databases are currently available for herbs.

Several databases were developed that contained traditional medicine-related data, such as "omics" data (Wang et al., 2018) and information on relationships between herbs and compounds or diseases (Fang et al., 2021; Ru et al., 2014; Xue et al., 2013). An integrated database that focuses on herb genomics and covers a wide range of traditional medicinal species is increasingly required for rapidly emerging herb genomic data. In this study, we constructed a Global Pharmacopoeia Genome Database (GPGD), which collected herb species genomic data from global pharmacopoeias to provide a mineable resource for traditional medicine research. The GPGD is an all-inclusive, reliable "omics" database incorporating several data mining functions. A set of convenient bioinformatics tools, such as BLAST (Mount, 2007) and JBrowse (Buels et al., 2016), were embedded in the system. The GPGD is publicly accessible at http://www.gpgenome.com/.

RESULTS

Genomic data stored in the GPGD

The GPGD includes four data fields; DNA barcodes, genomes (nuclear and organellar), individual genomic fragments (whole-genome sequencing at a low depth), and transcriptomes. In total, 34,346 records for 903 species belong-

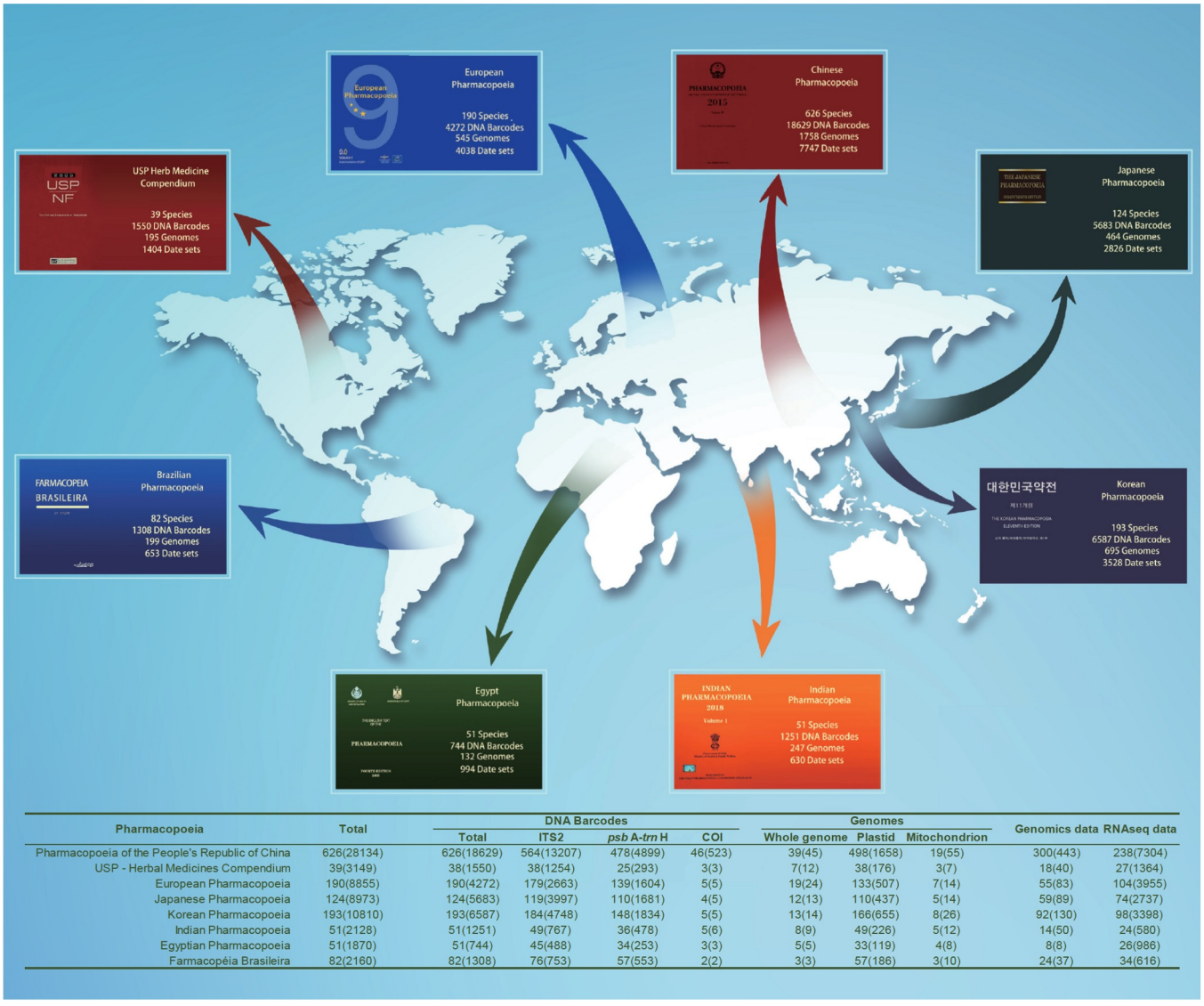
ing to 519 genera were collected from eight pharmacopoeias, including 21,872 DNA barcodes from 867 species, 2,203 organelle genomes from 674 species, 55 whole genomes from 49 species, and 9,682 transcriptome datasets from 350 species (Figure 1). In particular, 24.23% (534/2,203) of the plastid genomes of all species were newly generated in this study (Table S2 in Supporting Information). The plastid genomes of the genera *Cissampelos*, *Commelina*, and *Phyllanthus* were first characterized in this study. All data were organized and stored in a MySQL database or file storage server, and convenient data browsing and extraction tools were implemented in the GPGD Figures 2 and 3A).

Information browsing and searching

Multiple links, which connect to the home page, pharmacopoeia introduction, herb species information, information searches and tools, are listed at the frontend of the GPGD (Figure 3B). A brief description of each pharmacopoeia is included in the pharmacopoeia introduction page, including therein herb species. Taxonomy, pharmacopoeias, and genomic data download/visualization links are listed on the herb species information page for each herb. Information searches may be conducted using query keywords (such as species or genes) using the TNT search engine (https://github.com/teamtnt/tntsearch), and search results will include links to related data pages (Figure 3B–D). The tool page includes links to the following functions: species identification, genome visualization, sequence searching, and data retrieval.

Species identification

Currently, DNA barcodes (ITS2, *psb*A-*trn*H, and COI markers) and organelle genome sequences are used for herb



Pharmacopoeia	Total	DNA Barcodes				Genomes			Genomics data	RNAseq data
		Total	ITS2	*psb*A-*trn*H	COI	Whole genome	Plastid	Mitochondrion		
Pharmacopoeia of the People's Republic of China	626(28134)	626(18629)	564(13207)	478(4899)	46(523)	39(45)	498(1658)	19(55)	300(443)	238(7304)
USP - Herbal Medicines Compendium	39(3149)	38(1550)	38(1254)	25(293)	3(3)	7(12)	38(176)	3(7)	18(40)	27(1364)
European Pharmacopoeia	190(8855)	190(4272)	179(2663)	139(1604)	5(5)	19(24)	133(507)	7(14)	55(83)	104(3955)
Japanese Pharmacopoeia	124(8973)	124(5683)	119(3997)	110(1681)	4(5)	12(13)	110(437)	5(14)	59(89)	74(2737)
Korean Pharmacopoeia	193(10810)	193(6587)	184(4748)	148(1834)	5(5)	13(14)	166(655)	8(26)	92(130)	98(3398)
Indian Pharmacopoeia	51(2128)	51(1251)	49(767)	36(478)	5(6)	8(9)	49(226)	5(12)	14(50)	24(580)
Egyptian Pharmacopoeia	51(1870)	51(744)	45(488)	34(253)	3(3)	5(5)	33(119)	4(8)	8(8)	26(986)
Farmacopéia Brasileira	82(2160)	82(1308)	76(753)	57(553)	2(2)	3(3)	57(186)	3(10)	24(37)	34(616)

Figure 1 Information on herb species genomics data in GPGD.

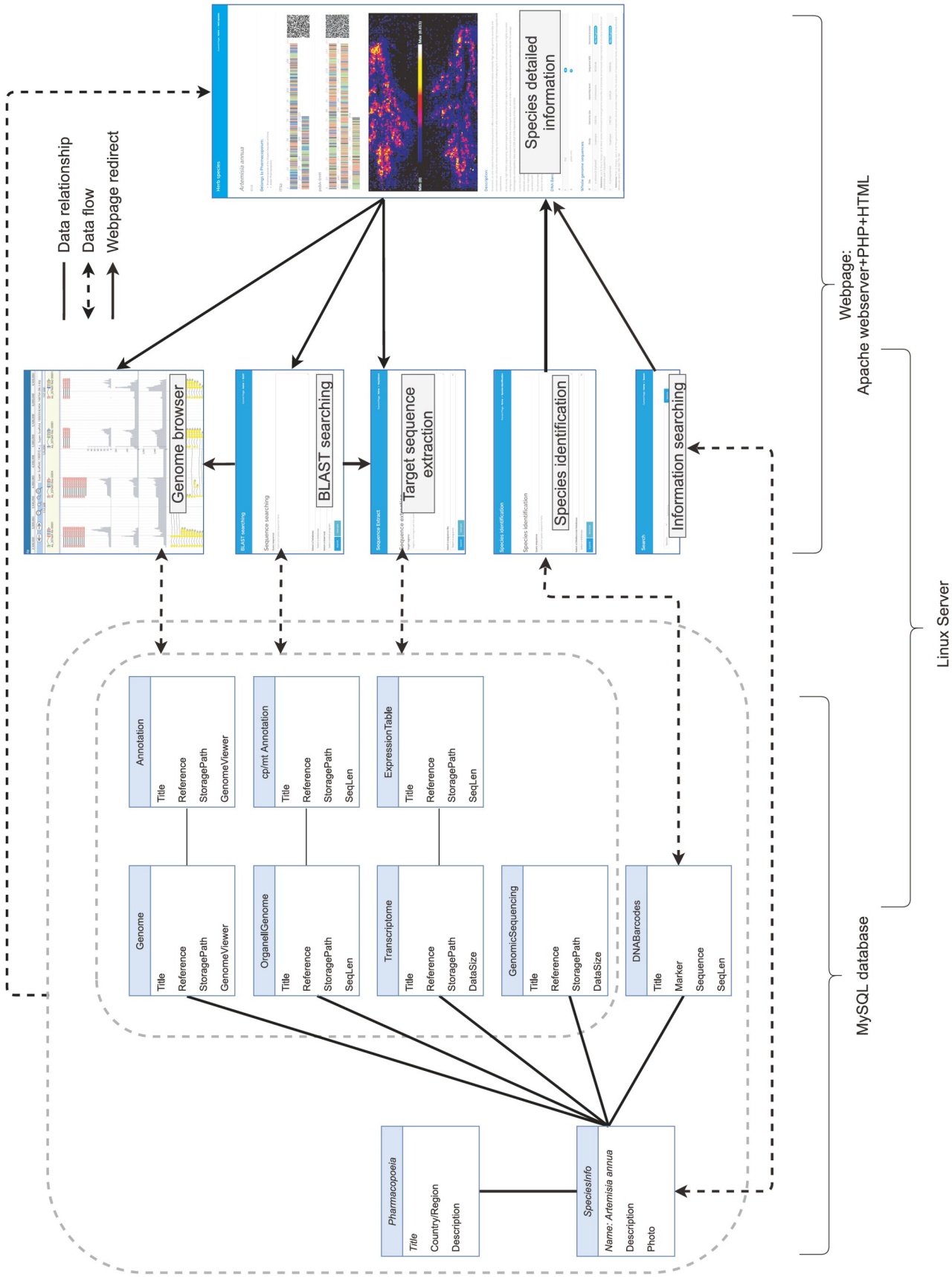


Figure 2 Database structure and GPGD webpages.

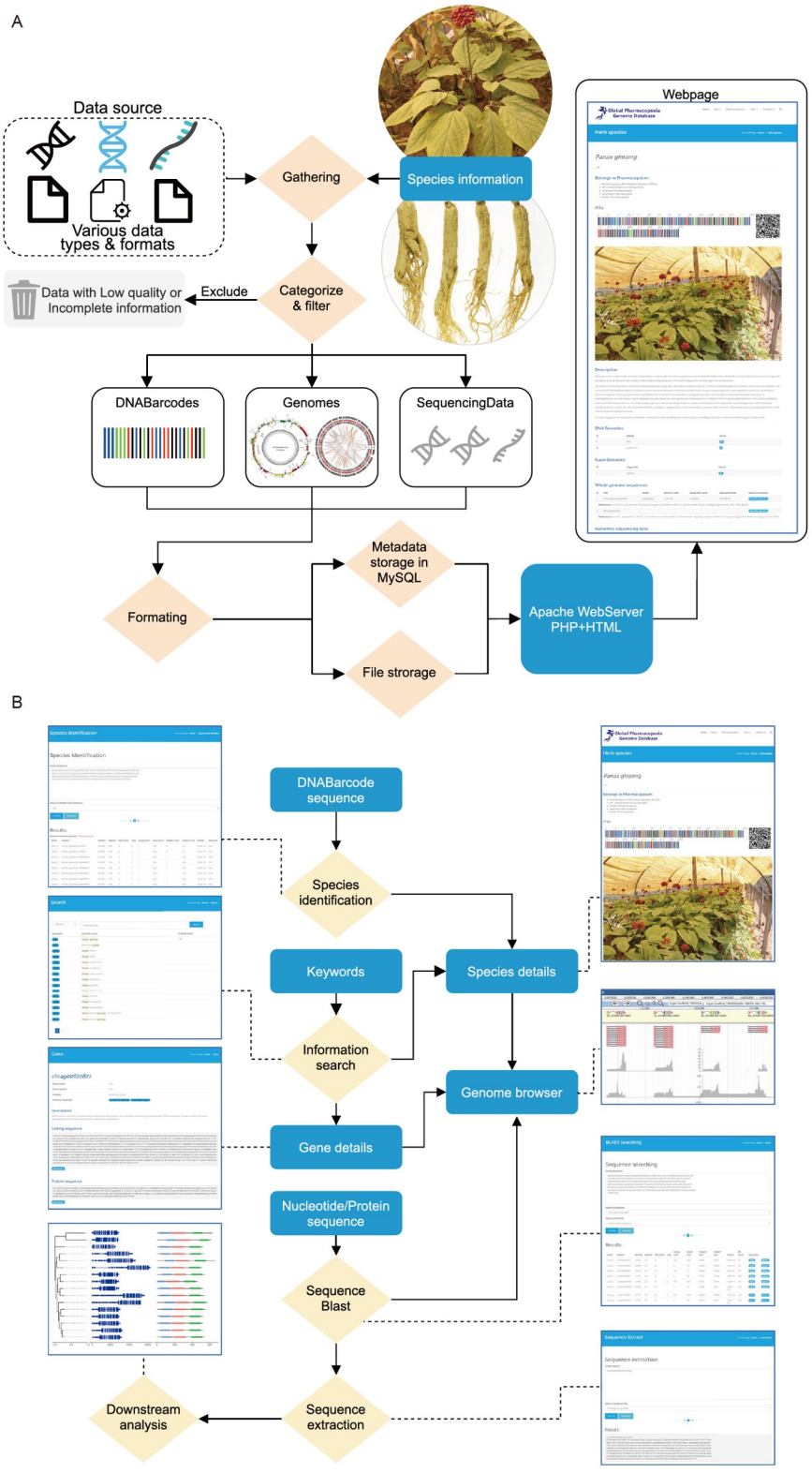


Figure 3 Data collection and usage workflow in the GPGD. A, Genomic data collection workflow. B, A usage workflow example: the GPGD provides information searching, species identification based on DNA barcodes, sequence searching and extraction from genome sequences, and data visualization in the genome browser.

species identification (Chen et al., 2010; Hebert et al., 2003; Li et al., 2015). In total, 22,686 DNA barcodes covering 1,223 species were collected in the GPGD (Table S2 in Supporting Information, http://www.gpgenome.com/spe-

cies). Species identification may be achieved by comparing submitted sequences with DNA barcodes or organelle genomes stored in the GPGD using BLASTn. Users input a sequence and select a sequence database from ITS2, *psb*A-*trn*H, COI, or plastid/mitochondrial genomes, and the best-hit species, as well as the top 50 hits, are returned.

Genome data visualization

Nuclear genome data were collected from publicly available genome projects (stored in the National Center for Biotechnology Information (NCBI) assembly database or independent databases from several research groups) and were newly generated by our group (Table S2 in Supporting Information). Organelle genomes were downloaded from public projects stored in the NCBI nucleotide database or generated in this study. Sequence searches may be performed with embedded BLAST tools (Mount, 2007). Genomic fragments which represent low coverage whole-genome sequencing data were generated in this study. BAM (read alignment) tracks and SNP tracks were generated if a whole-genome assembly was available. All genomes (nuclear and organellar) may be viewed and retrieved using JBrowse (Buels et al., 2016) (Figure 3).

Sequence searching and data retrieval

Users may retrieve genomic data from the GPGD for further mining by direct download or target sequence searching. BLAST databases were built for the sequences stored in the GPGD, allowing users to search all sequences in the GPGD by submitting a query sequence. Target regions may be extracted using the extraction tool.

DISCUSSION

Herbgenomics introduces cutting-edge "omics" technologies to traditional medicine research and enables the investigation of disease prevention and treatment mechanisms of traditional medicines (Hu et al., 2019). Considerable progress in sequencing technologies has promoted the rapid growth of herb genomic data, with multiple applications for various fields including, species identification, evolutionary history discovery, and gene function/biosynthetic pathway resolution. However, limitations such as non-uniform formats and storage at different locations have led to the inefficient use of traditional medicine genomic data.

In this study, a comprehensive GPGD for herb genomics data was constructed. Multiple layers of medicinal species genomic data were collected and uniformly formatted. The comprehensive genomic data stored in the GPGD can be used to identify medicinal species, study biosynthetic pathways of key secondary metabolites, and facilitate the genetic breeding of elite cultivars (Figure 4). Pharmacopoeias are legally binding collections of standards and quality specifications of medicinal products in the interest of public used in national market (World Health Organization, 2013). The GPGD is the first online system aimed at regulated herbs recorded in pharmacopoeias on the global market. Therefore, the utilization of "omics" data for the authentication, cultivation, and breeding of herbs sets the GPGD apart from common science- or interest-oriented medicinal species databases (Fang et al., 2021; Ru et al., 2014; Wang et al., 2018; Xue et al., 2013). The GPGD can be extensively used for herb research, thus contributing to global human health. Importantly, GPGD improvements are continuously ongoing. As herb genomic data emerges at accelerated rates, more species and pharmacopoeias will be incorporated, and more tools will be developed to meet the requirements of future diverse herb studies. In addition to species recorded in pharmacopoeias, genomic data from relative species and species having significant medicinal value will be included in the future. Importantly, these additions will include medicinal species (Lv et al., 2020; Vadakkemukadiyil Chellappan et al., 2019) used by certain nationalities or regions.

In summary, the GPGD provides a comprehensive resource of reliable "omics" data and frequently used bioinformatics analysis capabilities for herbal medicinal plants. We regularly update the database with newly added datasets and constantly improve the GPGD with enhanced functionalities to provide a more valuable resource to expedite species identification, decomposition of biosynthetic pathways, molecular-assisted breeding, and other related herb genomic studies. More data types (e.g., proteomics and metabolomics) and more functions (e.g., the identification of relationships between gene expression and metabolites) will be included or developed in the future.

MATERIALS AND METHODS

Organelle genome generation pipeline

Young tender leaves were collected from most medicinal species, and dried tissues were collected if fresh leaves were unavailable. DNA was isolated using the modified cetyltrimethylammonium bromide method (Li et al., 2013). Paired-end libraries with average insert sizes of 350 base pairs (bp) were constructed and sequenced on the Illumina HiSeq XTen platform (Illumina Inc., USA), and at least 4 Gb of sequencing data were generated for each sample. Organelle genomes were assembled using a reference-based approach. Reference genomes of plastids and mitochondria were downloaded from NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/). Sequencing reads were mapped to reference genomes, and mapped reads were extracted for plastids and mitochondria.

Extracted reads were assembled into contigs using ABySS 2.0 (Jackman et al., 2017). The contigs were ordered and connected manually according to the most closely related reference sequence. Owing to their high variability and because few plant mitochondrial genomes were available, only a few (37 samples) mitochondrial genomes were successively assembled. The plastid genome was annotated using Plann (Huang and Cronk, 2015), and the mitochondrial genome was annotated using MitoZ (Meng et al., 2019).

Whole-genome gene prediction and annotation

Gene annotations of the whole genome were included with published genome projects, if available. Gene structure predictions and functional annotations were performed for genomes with unavailable gene model information. Gene predictions were conducted using the Maker-P pipeline (Campbell et al., 2014) with evidences of assembled transcripts from RNA-seq data or transcripts from closely related species and protein sequences from UniProt (https://www.uniprot.org/). Gene function annotations were conducted using InterProScan v5.26 (Jones et al., 2014) and PfamScan (El-Gebali et al., 2018). All gene annotation results were formatted in the GFF3 format. Transcription factors were predicted by combining PfamScan results and a BLAST search of PlantTFDB (Jin et al., 2016). Disease resistance (*R*) genes were predicted using PRGdb 3.0 (Osuna-Cruz et al., 2017).

Gene information may be searched using the search page. In addition, gene sequences may be searched using BLAST tools. Transcriptome data were collected from the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra). The GPGD stores the metadata and gene expression profiles calculated based on each dataset using HISAT2 and StringTie (Pertea et al., 2016). Gene information may be viewed from the JBrowse genome page (Figure 4).

Data collection and manual curation workflow

Herb species genomic data were collected from public databases and also newly generated in this study. All data were categorized into four fields: DNA barcodes, super-barcodes, genomes, and sequencing datasets. Each data type was manually inspected and converted to the same format. For DNA barcodes, sequences having the best hits belonging to different species in self-vs.-self BLAST searches will be filtered out. For super-barcodes, assembled contigs with total lengths <20 kb will be removed. For genomes: (i) genomes from species that are also common crops (e.g., rice and soybean) will not be included in the database; (ii) the latest version of the genome will be kept if multiple versions exist

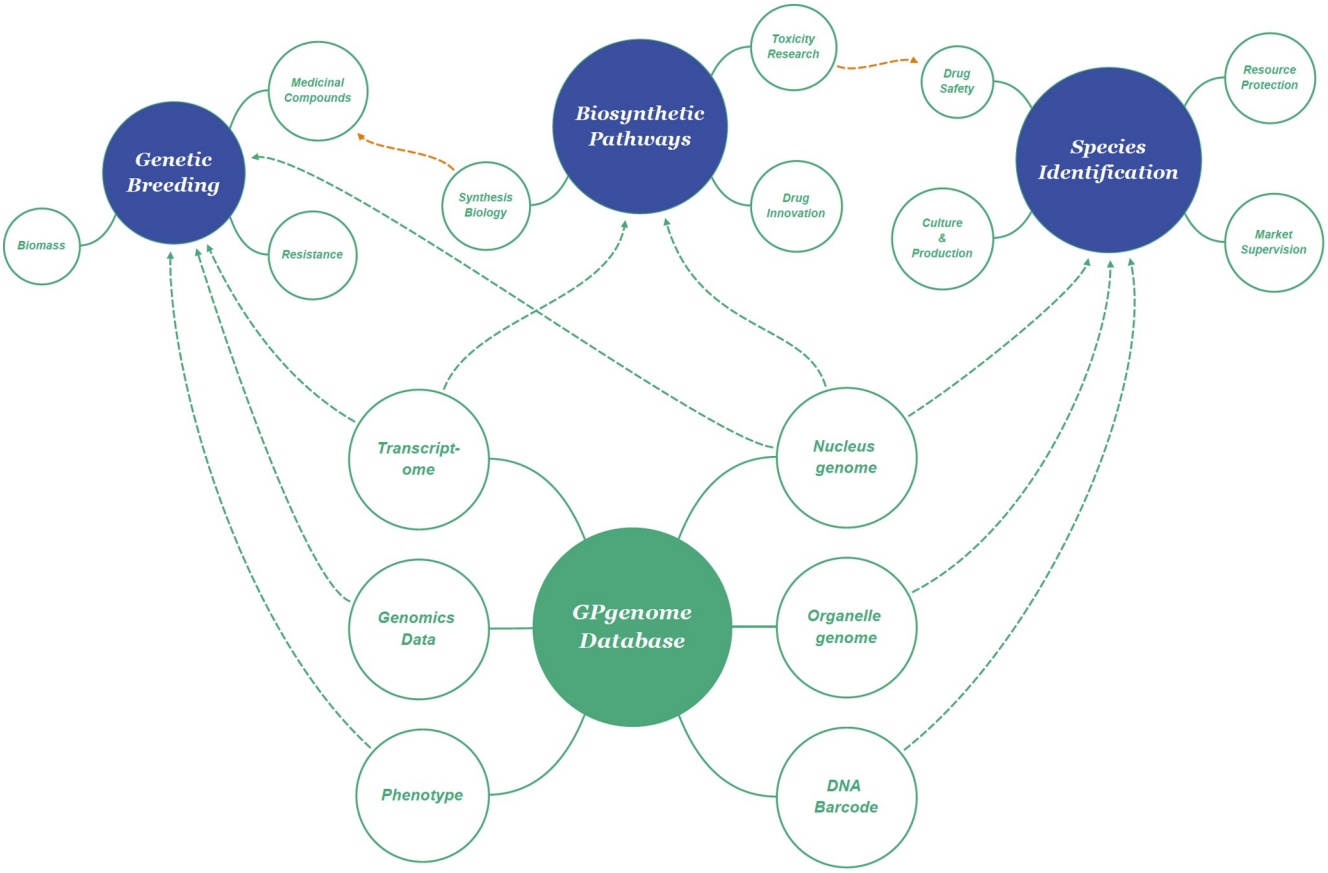


Figure 4 Application fields for the GPGD.

and come from the same strain; and (iii) all versions will be kept if they come from different strains. For sequencing datasets from next-generation sequencing, low-quality data with a base Q20<95% will be removed, and no filtering for sequencing datasets from third-generation sequencing will be performed at present.

Implementation and database access

Metadata from the four fields, DNA barcodes, genomes (nuclear and organellar), individual genomic fragments (whole-genome sequencing at a low depth), and transcriptomes were organized in a MySQL database with species information as a core table (Figure 2). Public and newly generated data were filtered and processed into a uniform format (Figure 3A). For instance, all genome annotations were stored as zipped gff files, and gene expression information was stored as counts and transcripts per million tables. Data may be accessed via the web server; http://www.gpgenome.com/. The web server was built based on LAMP (Linux+Apache+MySQL+PHP), and the webpage was constructed on a PHP MVC framework (Laravel v5.8; https://laravel.com/) (Figure 3A).

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *This work was supported by the National Key Research and Development Program of China (2019YFC1711100), the National Natural Science Foundation of China and Karst Science Research Center of Guizhou Province (U1812403-1), the Special Foundation for National Science and Technology Basic Research Program of China (2018FY100701), the Fundamental Research Funds for the Central Public Welfare Research Institutes (ZXKT17027, ZXKT18014), the Open Research Fund of Chengdu University of Traditional Chinese Medicine Key Laboratory of Systematic Research of Distinctive Chinese Medicine Resources in Southwest China (2020GZ2011016), the Funds for Fostering Outstanding Scholars in Science and Technology (Innovation) (ZZ13-YQ-047), and Innovation Fund of China Academy of Chinese Medical Sciences. The authors thank Professor Salwa El-Hallouty and Professor Rehab Ahmed (Department of Pharmacognosy, National Research Center, Egypt) for providing medicinal species information from the Egyptian Pharmacopoeia. We also thank Professor Baozhong Duan (College of Pharmaceutical Science, Dali University), Professor Dianyun Hou (Agricultural College, Henan University of Science and Technology), Dr. Qinghua Wu (Chengdu University of Traditional Chinese Medicine), Professor Xiaojun Zhang (Heilongjiang Mudanjiang Normal University), Dr. Weisi Ma, and Fuhe Chen for providing medicinal species samples.*

References

Alaux, M., Rogers, J., Letellier, T., Flores, R., Alfama, F., Pommier, C., Mohellibi, N., Durand, S., Kimmel, E., Michotey, C., et al. (2018). Linking the international Wheat Genome Sequencing Consortium bread wheat reference genome sequence to wheat genetic and phenomic data. Genome Biol 19, 111.

Atanasov, A.G., Waltenberger, B., Pferschy-Wenzig, E.M., Linder, T., Wawrosch, C., Uhrin, P., Temml, V., Wang, L., Schwaiger, S., Heiss, E. H., et al. (2015). Discovery and resupply of pharmacologically active plant-derived natural products: A review. Biotech Adv 33, 1582–1614.

Bantawa, P., Ghosh, S.K., Maitra, S., Ghosh, P.D., and Mondal, T.K. (2009). Status and conservation threats of *Picrorhiza scrophulariiflora* Pennell. (Scrophulariaceae): An endangered high valued medicinal plant of Indo-China Himalayan region. Bioremed Biodivers Bioavailab 3, 15–22.

Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M., Elsik, C.G., Lewis, S.E., Stein, L., et al. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol 17, 66.

Callaway, E., and Cyranoski, D. (2015). Anti-parasite drugs sweep Nobel prize in medicine 2015. Nature 526, 174–175.

Campbell, M.S., Law, M.Y., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., et al. (2014). MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol 164, 513–524.

Chan, T.Y.K. (2002). Incidence of herb-induced aconitine poisoning in Hong Kong. Drug Saf 25, 823–828.

Chen, S., Pang, X., Song, J., Shi, L., Yao, H., Han, J., and Leon, C. (2014). A renaissance in herbal medicine identification: from morphology to DNA. Biotech Adv 32, 1237–1244.

Chen, S., Su, G., Zou, J., Huang, L., Guo, B., and Xiao, P. (2005). The sustainable development framework of national Chinese medicine resources (in Chinese). China J Chin Mater Med 30, 1141–1146.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., et al. (2010). Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS ONE 5, e8613.

Chen, X., Zhou, J., Cui, Y., Wang, Y., Duan, B., and Yao, H. (2018). Identification of *Ligularia* Herbs using the complete chloroplast genome as a super-barcode. Front Pharmacol 9, 695.

Chik, W.I., Zhu, L., Fan, L.L., Yi, T., Zhu, G.Y., Gou, X.J., Tang, Y.N., Xu, J., Yeung, W.P., Zhao, Z.Z., et al. (2015). *Saussurea involucrata*: A review of the botany, phytochemistry and ethnopharmacology of a rare traditional herbal medicine. J Ethnopharmacol 172, 44–60.

Coordinating Research Group for the Structure of Artemisinin. (1977). Artemisinin: A new type of sesquiterpene lactone (in Chinese). Chin Sci Bull 22, 142.

Cordell, G.A. (2015). Ecopharmacognosy and the responsibilities of natural product research to sustainability. Phytochem Lett 11, 332–346.

Dong, L., Chen, Z., Wang, Y., Wei, F., Zhang, L., Xu, J., Wei, G., Wang, R., Yang, J., and Liu, W. (2017). DNA marker-assisted selection of medicinal plants (I). Breeding research of disease-resistant cultivars of *Panax notoginseng* (in Chinese). China J Chin Mater Med 42, 56–62.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2018). The Pfam protein families database in 2019. Nucleic Acids Res 47, D427–D432.

Fang, S.S., Dong, L., Liu, L., Guo, J.C., Zhao, L.H., Zhang, J.Y., Bu, D.C., Liu, X.K., Huo, P.P., Cao, W.C., et al. (2021). HERB: a high-throughput experiment- and reference-guided database of traditional Chinese medicine. Nucleic Acids Res 49, D1197–D1206.

Geuna, F. (2017). Molecular-assisted breeding. In: Pilu, R., and Gavazzi, G., eds. More Food: Road to Survial. Sharjah: Bentham Science Publishers. 373–398.

Gilbert, N. (2011). Regulations: Herbal medicine rule book. Nature 480, S98–S99.

Guan, Z., Li, Z., and Li, E. (2002). A rare and endangered medicinal plant: *Dendrobium nobile* (in Chinese). Chin Wild Plant Resour 21, 36–37.

Guo, L., Winzer, T., Yang, X., Li, Y., Ning, Z., He, Z., Teodor, R., Lu, Y., Bowser, T.A., Graham, I.A., et al. (2018). The opium poppy genome and morphinan production. Science 362, 343–347.

Guo, M., Ren, L., Xu, Y., Liao, B., Song, J., Li, Y., Mantri, N., Guo, B., Chen, S., and Pang, X. (2019). Development of plastid genomic resources for discrimination and classification of *Epimedium wushanense* (Berberidaceae). Int J Mol Sci 20, 4003.

Han, J., Pang, X., Liao, B., Yao, H., Song, J., and Chen, S. (2016). An authenticity survey of herbal medicines from markets in China using

DNA barcoding. Sci Rep 6, 18723.
Hebert, P.D.N., Cywinska, A., Ball, S.L., and deWaard, J.R. (2003). Biological identifications through DNA barcodes. Proc R Soc Lond B 270, 313–321.
Hu, H., Shen, X., Liao, B., Luo, L., Xu, J., and Chen, S. (2019). Herbgenomics: A stepping stone for research into herbal medicine. Sci China Life Sci 62, 913–920.
Huang, D.I., and Cronk, Q.C.B. (2015). Plann: A command-line application for annotating plastome sequences. Appl Plant Sci 3, 1500026.
Jackman, S.D., Vandervalk, B.P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S.A., Jahesh, G., Khan, H., Coombe, L., Warren, R.L., et al. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res 27, 768–777.
Jin, J., Tian, F., Yang, D.C., Meng, Y.Q., Kong, L., Luo, J., and Gao, G. (2016). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res 45, D1040–D1045.
Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240.
Kim, N.H., Jayakodi, M., Lee, S.C., Choi, B.S., Jang, W., Lee, J., Kim, H. H., Waminal, N.E., Lakshmanan, M., van Nguyen, B., et al. (2018). Genome and evolution of the shade-requiring medicinal herb *Panax ginseng*. Plant Biotechnol J 16, 1904–1917.
Li, J., Wang, S., Yu, J., Wang, L., and Zhou, S. (2013). A modified CTAB protocol for plant DNA extraction. Chin Bull Bot 48, 72–78.
Li, X., Yang, Y., Henry, R.J., Rossetto, M., Wang, Y., and Chen, S. (2015). Plant DNA barcoding: from gene to genome. Biol Rev 90, 157–166.
Liang, H.H., Cheng, Z., Yang, X.L., Li, S., Ding, Z.Q., Zhou, T.S., Zhang, W.J., and Chen, J.K. (2008). Genetic diversity and structure of *Cordyceps sinensis* populations from extensive geographical regions in China as revealed by inter-simple sequence repeat markers. J Microbiol 46, 549–556.
Lowe, L., Matteucci, M.J., and Schneir, A.B. (2005). Herbal aconite tea and refractory ventricular tachycardia. N Engl J Med 353, 1532.
Lu, J., and Lan, X. (2013). An investigation on rare and endangered Tibetan medicinal plants in Lhasa region (in Chinese). China J Chin Mater Med 38, 127–132.
Luo, L., Jiang, J., Wang, C., Fitzgerald, M., Hu, W., Zhou, Y., Zhang, H., and Chen, S. (2020). Analysis on herbal medicines utilized for treatment of COVID-19. Acta Pharm Sin B 10, 1192–1204.
Lv, Q., Qiu, J., Liu, J., Li, Z., Zhang, W., Wang, Q., Fang, J., Pan, J., Chen, Z., Cheng, W., et al. (2020). The *Chimonanthus salicifolius* genome provides insight into magnoliid evolution and flavonoid biosynthesis. Plant J 103, 1910–1923.
Meng, G., Li, Y., Yang, C., and Liu, S. (2019). MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic Acids Res 47, e63.
Mount, D.W. (2007). Using the basic local alignment search tool (BLAST). Cold Spring Harb Protoc 2007(7), pdb.top17.
Muniraj, N., Siddharth, S., and Sharma, D. (2019). Bioactive compounds: Multi-targeting silver bullets for preventing and treating breast cancer. Cancers 11, 1563.
Newman, D.J., and Cragg, G.M. (2020). Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. J Nat Prod 83, 770–803.
Osuna-Cruz, C.M., Paytuvi-Gallart, A., Di Donato, A., Sundesha, V., Andolfo, G., Aiese Cigliano, R., Sanseverino, W., and Ercolano, M.R. (2017). PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes. Nucleic Acids Res 46, D1197–D1201.
Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., and Salzberg, S.L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 11, 1650–1667.
Pešić, M. (2015). Development of natural product drugs in a sustainable manner. In: Brief for United Nations Global Sustainable Development Report 2015.
Poon, W.T., Lai, C.K., Ching, C.K., Tse, K.Y., So, Y.C., Chan, Y.C., Hau, L. M., Mak, T.W.L., and Chan, A.Y.W. (2006). Aconite poisoning in camouflage. Hong Kong Med J 12, 456–459.
Portwood, J.L., Woodhouse, M.R., Cannon, E.K., Gardiner, J.M., Harper, L.C., Schaeffer, M.L., Walsh, J.R., Sen, T.Z., Cho, K.T., Schott, D.A., et al. (2019). MaizeGDB 2018: the maize multi-genome genetics and genomics database. Nucleic Acids Res 47, D1146–D1154.
Ru, J., Li, P., Wang, J., Zhou, W., Li, B., Huang, C., Li, P., Guo, Z., Tao, W., Yang, Y., et al. (2014). TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. J Cheminform 6, 1–6.
Shen, Q., Zhang, D., Sun, W., Zhang, Y., Shang, Z., and Chen, S. (2017a). Medicinal plant DNA marker assisted breeding (II) the assistant identification of SNPs assisted identification and breeding research of high yield *Perilla frutescens* new variety (in Chinese). China J Chin Mater Med 42, 1668–1672.
Shen, X., Wu, M., Liao, B., Liu, Z., Bai, R., Xiao, S., Li, X., Zhang, B., Xu, J., and Chen, S. (2017b). Complete chloroplast genome sequence and phylogenetic analysis of the medicinal plant *Artemisia annua*. Molecules 22, 1330.
Smanski, M.J., Zhou, H., Claesen, J., Shen, B., Fischbach, M.A., and Voigt, C.A. (2016). Synthetic biology to access and expand nature's chemical diversity. Nat Rev Microbiol 14, 135–149.
Tareke Woldegiorgis, S., Wang, S., He, Y., Xu, Z., Chen, L., Tao, H., Zhang, Y., Zou, Y., Harrison, A., Zhang, L., et al. (2019). Rice stress-resistant SNP database. Rice 12, 97.
Vadakkemukadiyil Chellappan, B., Pr, S., Vijayan, S., Rajan, V.S., Sasi, A., and Nair, A.S. (2019). High quality draft genome of Arogyapacha (*Trichopus zeylanicus*), an important medicinal plant endemic to Western Ghats of India. G3 9, 2395–2404.
Wang, X., Zhang, J., He, S., Gao, Y., Ma, X., Gao, Y., Zhang, G., Kui, L., Wang, W., Wang, Y., et al. (2018). HMOD: an omics database for herbal medicine plants. Mol Plant 11, 757–759.
Weaver, B.A. (2014). How Taxol/paclitaxel kills cancer cells. Mol Bio Cell 25, 2677–2681.
World Health Organization. (2013). The international pharmacopoeia. WHO Drug Inf 27, 119–128.
World Health Organization. (2019). Index of Pharmacopoeias. Geneva: World Health Organization.
Xiao, P., Zhao, R., Long, X., and Guo, B. (2009). Macroscopic analysis on production and marketing of medicinal material resources for sustainable development (in Chinese). China J Chin Mater Med 34, 2135–2139.
Xu, J., Chu, Y., Liao, B., Xiao, S., Yin, Q., Bai, R., Su, H., Dong, L., Li, X., Qian, J., et al. (2017). *Panax ginseng* genome examination for ginsenoside biosynthesis. Gigascience 6, gix093.
Xue, R., Fang, Z., Zhang, M., Yi, Z., Wen, C., and Shi, T. (2013). TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis. Nucleic Acids Res 41, D1089–D1095.

SUPPORTING INFORMATION

The supporting information is available online at https://doi.org/10.1007/s11427-021-1968-7. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.