OXFORD

Briefings in Bioinformatics, 21(6), 2020, 1857–1874

doi: 10.1093/bib/bbz132
Advance Access Publication Date: 18 November 2019
Problem solving protocol

Bioinformatics-assisted, integrated omics studies on medicinal plants

Xiaoxia Ma , Yijun Meng, Pu Wang, Zhonghai Tang, Huizhong Wang and Tian Xie

Corresponding authors: Yijun Meng: Hangzhou Normal University, Hangzhou 311121, P.R. China. Tel: 86-571-28865198. Fax: 86-571-28865198. E-mail: mengyijun@zju.edu.cn; Tian Xie: Hangzhou Normal University, Hangzhou 311121, P.R. China. Tel: 86-571-28860237. Fax: 86-571-28860237. E-mail: xbs@hznu.edu.cn

Abstract

The immense therapeutic and economic values of medicinal plants have attracted increasing attention from the worldwide researchers. It has been recognized that production of the authentic and high-quality herbal drugs became the prerequisite for maintaining the healthy development of the traditional medicine industry. To this end, intensive research efforts have been devoted to the basic studies, in order to pave a way for standardized authentication of the plant materials, and bioengineering of the metabolic pathways in the medicinal plants. In this paper, the recent advances of omics studies on the medicinal plants were summarized from several aspects, including phenomics and taxonomics, genomics, transcriptomics, proteomics and metabolomics. We proposed a multi-omics data-based workflow for medicinal plant research. It was emphasized that integration of the omics data was important for plant authentication and mechanistic studies on plant metabolism. Additionally, the computational tools for proper storage, efficient processing and high-throughput analyses of the omics data have been introduced into the workflow. According to the workflow, authentication of the medicinal plant materials should not only be performed at the phenomics level but also be implemented by genomic and metabolomic marker-based examination. On the other hand, functional genomics studies, transcriptional regulatory networks and protein–protein interactions will contribute greatly for deciphering the secondary metabolic pathways. Finally, we hope that our work could inspire further efforts on the bioinformatics-assisted, integrated omics studies on the medicinal plants.

Key words: integrated omics studies; medicinal plants; databases; software; authentication; secondary metabolism

Introduction

Medicinal plants constitute a huge library of natural organic compounds with promising pharmaceutical application prospects. As early as in ancient China, diverse plant species had been used as the crude materials for preparation of traditional medicines. With the development of modern medicine, a significant portion of the commercially available drugs were purified or

Xiaoxia Ma is a research assistant in Hangzhou Normal University. Currently, she is focusing on the transcriptomics and metabolomics studies on medicinal plants.
Yijun Meng is a professor in Hangzhou Normal University. His current research interests are mainly focusing on the discovery and functional characterization of the non-coding RNAs in plants. He is one of the corresponding authors of this paper.
Pu Wang is a professor in Zhejiang University of Technology. Her current research interest is mainly focusing on microbiology-based biomedical chemistry.
Zhonghai Tang is an associate professor in Hunan Agricultural University. His major research topics include identification of the bioactive integrates from the herbal plants and deciphering the secondary metabolic pathways in medicinal plants.
Huizhong Wang is a professor in Hangzhou Normal University. He is currently working on elucidating the molecular mechanisms of secondary metabolic pathways of the valuable Chinese medicinal herbs.
Tian Xie is a professor in Hangzhou Normal University. He is currently working on the development of novel anti-cancer drugs from the medicinal plants. He is one of the corresponding authors of this paper.
Submitted: 5 July 2019; **Received (in revised form):** 3 September 2019

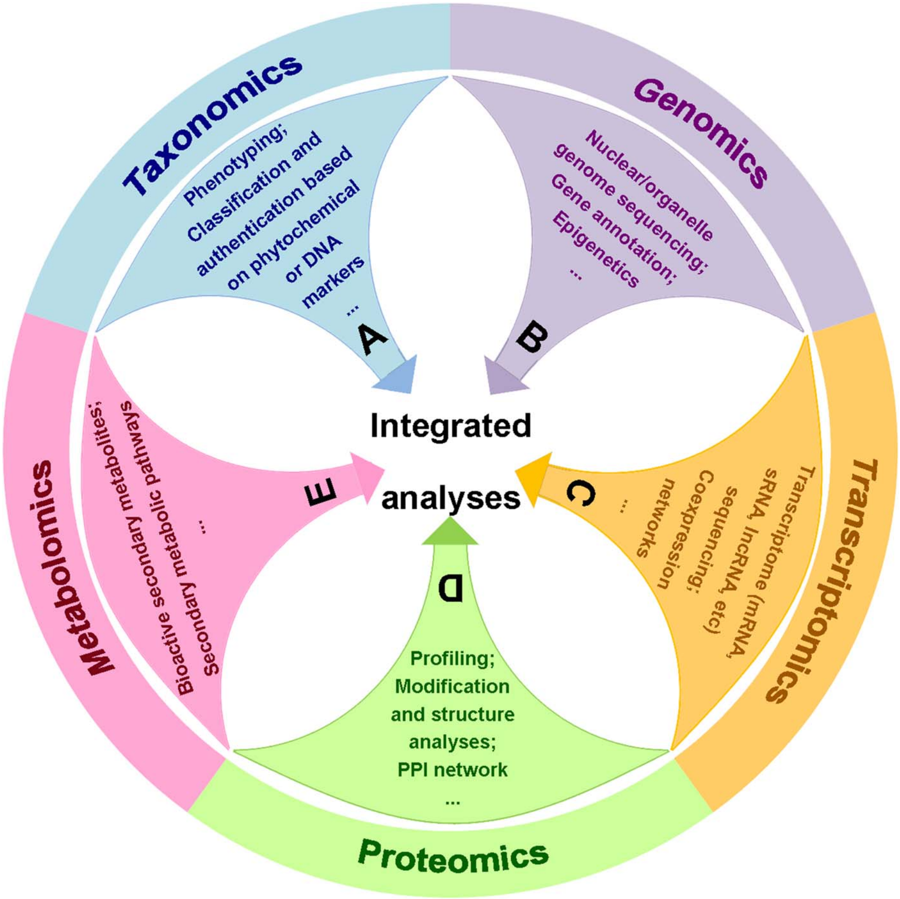

Figure 1. Summary of the multi-omics data-based integrated studies on medicinal plants. (A) Phenomics and taxonomics studies, including (1) phenotyping of the medicinal plants growing under distinct conditions or at different stages and (2) classification and authentication based on DNA and/or phytochemical markers. (B) Genomics studies, including (1) nuclear and organelle genome sequencing or resequencing, (2) gene annotation and (3) epigenetics analysis. (C) Transcriptomics studies, including (1) transcriptome sequencing for gene expression profiling, (2) co-expression network construction and (3) non-coding RNA discovery. (D) Proteomics studies, including (1) protein identification and quantification, (2) protein modification and structure analyses and (3) protein–protein interaction network construction. (E) Metabolomics studies, including (1) qualitative and quantitative analyses of secondary metabolites, (2) deciphering regulatory mechanisms underlying secondary metabolic pathways and (3) biotarget discovery and medicinal value exploration.

modified from the plant-originated natural ingredients [1]. To date, the medicinal plants have brought great benefits to the pharmaceutical industry all over the world. At the same time, modernized and standardized research on the medicinal plants becomes a pressing issue to maintain the healthy and sustainable growth of the biomedicine industry. There are two major tasks related to the medicinal plant research, including authentication of the plant materials and mechanistic studies on the metabolic pathways. To accomplish the above tasks, varied omics technologies, such as phenotyping, genome sequencing, transcriptome sequencing and proteome profiling, have been widely applied for the medicinal plant research [2]. With the advances of single-cell sequencing technology, it is foreseeable that the plant metabolic pathways could be analyzed at the single-cell resolution [3]. On the other hand, systems biology was defined as a multi-disciplinary approach for decoding the complexity of biological systems. It requires joint efforts from biologists, chemists, mathematicians, physicists, and engineers for integrated analysis of different types of huge omics data sets. To date, several recent excellent reviews have emphasized the crucial role of systems biology approaches in deciphering gene regulatory networks and metabolic pathways in plants [4–6].

In this paper, we will summarize the state of the art of the omics studies on medicinal plants from five aspects, including phenomics and taxonomics, genomics, transcriptomics, proteomics and metabolomics (Figure 1). For instance, with the efforts on the genomics studies, the draft genomes of several medicinal plants have been released [7–14], which has greatly facilitated high-density DNA marker development, gene annotation and functional genomics studies. The next-generation sequencing (NGS) technology has been widely applied for the transcriptomics studies. NGS-based transcriptome sequencing enabled us to investigate the spatio-temporal expression patterns of specific genes [15, 16], to establish networks of the co-expressed genes [17, 18] and to discover the non-coding RNA (ncRNA) species in the medicinal plants [19–21]. Meanwhile, proper storage, efficient processing and integrated analysis of the growing omics data sets heavily rely on the availability of the powerful computational tools. In this regard, the bioinformatics databases and the software packages applicable to the medicinal plant research have been gathered in two reference lists (Table 1 and Table 2), respectively. For example, several databases have been constructed for accommodating omics data of the medicinal plants, such as phenotypic features, bioactive ingredients and their molecular targets, pharmacological uses and genome

Table 1. Non-exhaustive list of databases providing valuable omics resources for medicinal plant research

Classification	Database name	Brief description	URL	Citation
Phenomics and taxonomics	MPDB 1.0[a]	Medicinal plant database of Bangladesh	http://www.medicinalplantbd.net/ (valid)	[22]
	MPID[a]	Medicinal plant images database	https://library.hkbu.edu.hk/electronic/libdbs/mpd/ (valid)	
	ebDB[a]	The International Ethnobotany Database, a noncommercial repository for ethnobotanical data supporting multilingual functionality	https://ebdb.org/ (invalid)	
	PlantCLEF 2019[a]	Image-based identification of plant species	https://www.imageclef.org/PlantCLEF2019 (valid)	[23]
Genomics	CmMDb[a]	Versatile database for *Cucumis melo* (with economical and medicinal importance) microsatellite markers and other horticulture crop research	http://65.181.125.102/cmmdb2/index.html (invalid)	[24]
	HopBase[a]	Unified resource for Humulus genomics	http://hopbase.org/ (valid)	[25]
	MMDBD[a]	Medicinal materials DNA barcode database	http://www.cuhk.edu.hk/icm/mmdbd.htm (valid)	[26, 27]
	MGH[a]	Genome hub for the medicinal plant maca (*Lepidium meyenii*)	http://maca.eplant.org (valid)	[28]
Transcriptomics	DsTRD[a]	Danshen transcriptional resource database	http://bi.sky.zstu.edu.cn/DsTRD/home.php (invalid)	[29]
	GarlicESTdb[a]	Online database and mining tool for garlic EST sequences	http://garlicdb.kribb.re.kr (invalid)	[30]
	miRBase[b]	Providing basic information of miRNA genes from diverse species including some medicinal plants	http://www.mirbase.org/ (valid)	[31]
	MepmiRDB[b]	MiRNA database for medicinal plants.	http://mepmirdb.cn/mepmirdb/index.html (valid)	[32]
	ArrayExpress[a]	A public repository for microarray gene expression data from EBI	https://www.ebi.ac.uk/arrayexpress/ (valid)	[33]
	GEO[a]	Gene Expression Omnibus, a repository of high-throughput sequencing data from NCBI	https://www.ncbi.nlm.nih.gov/geo/ (valid)	[34]
	SRA[a]	Sequence Read Archive, a repository of high-throughput sequencing data from NCBI	https://www.ncbi.nlm.nih.gov/sra/ (valid)	[35]
	croFGD[b]	*Catharanthus roseus* functional genomics database	http://bioinformatics.cau.edu.cn/croFGD/ (valid)	[36]
Proteomics	UniProt[b]	Providing the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information	https://www.uniprot.org/ (valid)	[37]
	Pfam[b]	Large collection of protein families	http://pfam.xfam.org/ (valid)	[38]
	BioGRID[b]	Database dedicated to the curation and archival storage of protein, genetic and chemical interactions for all major model organism species and humans	https://thebiogrid.org/ (valid)	[39]
	STRING[b]	Protein–protein interaction networks	http://string-db.org (valid)	[40]
	IntAct[b]	Providing a freely available, open source database system and analysis tools for molecular interaction data	https://www.ebi.ac.uk/intact/ (valid)	[41]
	PAIR[b]	Predicted protein interactome resource of *Arabidopsis thaliana*	http://www.cls.zju.edu.cn/pair/ (invalid)	[42]
	PRIN[b]	Predicted protein interactome resource of *Oryza sativa*	http://bis.zju.edu.cn/prin/ (valid)	[43]
Metabolomics	CathaCyc[b]	Metabolic pathway database built from *Catharanthus roseus* RNA-Seq data	http://www.cathacyc.org (valid)	[44]
	KEGG PATHWAY Database[b]	Collection of manually drawn pathway maps representing the knowledge on the molecular interactions, reactions and relation networks such as for metabolism	https://www.kegg.jp/kegg/pathway.html (valid)	[45]

(Continued)

Table 1. Continued

Classification	Database name	Brief description	URL	Citation
Integrated omics	MPGR[a]	Medicinal Plants Genomics Resource, providing genomic, transcriptomic and metabolomic information of 14 medicinal plant species	http://medicinalplantgenomics.msu.edu/ (valid)	
	HMOD[a]	Omics database for herbal medicine plants	http://herbalplant.ynau.edu.cn/ (valid)	[46]
Biotarget discovery and medicinal value exploration	AromaDb[b]	Database of medicinal and aromatic plant's aroma molecules with phytochemistry and therapeutic potentials	http://bioinfo.cimap.res.in/aromadb/ (valid)	[47]
	Phytochemica[b]	Platform to explore phytochemicals of medicinal plants	http://home.iitj.ac.in/~bagler/webservers/Phytochemica (invalid)	[48]
	SerpentinaDB[b]	Database of plant-derived molecules of *Rauvolfia serpentine*	http://home.iitj.ac.in/&#x007E;bagler/webservers/SerpentinaDB/ (invalid)	[49]
	TCM Database@Taiwan[b]	Database containing free 3D molecular structure database of traditional Chinese medicine (TCM) available for virtual screening or molecular simulation.	http://tcm.cmu.edu.tw/ (valid)	[50]
	IMPPAT[b]	Curated database of Indian medicinal plants, providing 27,074 plant-phytochemical associations and 11,514 plant-therapeutic associations.	https://cb.imsc.res.in/imppat (site under maintenance)	[51]
	InDiaMed[b]	Comprehensive database of Indian medicinal plants for diabetes.	http://www.indiamed.info (invalid)	[52]
	MAPS[b]	Medicinal plant activities, phytochemical and structural database.	http://www.mapsdatabase.com (invalid)	[53]
	NeMedPlant[b]	Database of therapeutic applications and chemical constituents of medicinal plants from north-east region of India.	http://bif.uohyd.ac.in/nemedplant/ (valid)	[54]
	NPASS[b]	Natural product activity and species source database for natural product research, discovery and tool development.	http://bidd2.nus.edu.sg/NPASS/ (valid)	[55]
	CMAUP[b]	Database of collective molecular activities of useful plants, providing target information of the plant active ingredients.	http://bidd2.nus.edu.sg/CMAUP/ (valid)	[56]
	DIACAN[b]	Integrated database for antidiabetic and anticancer medicinal plants.	http://kaubic.in/diacan (invalid)	[57]
	SACPD[b]	Saudi anti-human cancer plants database.	https://teeqrani1.wixsite.com/sapd (valid)	[58]
	TarNet[b]	Manually curated database and platform of traditional medicinal plants with natural compounds that includes potential bio-target information.	http://www.herbbol.org:8001/tarnet (invalid)	[59]
	KNApSAcK[b]	Metabolite activity database comprising 9584 triplet relationships (metabolite–biological activity–target species).	http://kanaya.naist.jp/MetaboliteActivity/top.jsp (valid)	[60]
	TCMGeneDIT[b]	Database for associated TCM, gene and disease information using text mining	http://tcm.lifescience.ntu.edu.tw/ (valid)	[61]
	HIT[b]	Comprehensive and fully curated database for herb ingredients' targets	http://lifecenter.sgst.cn/hit/ (invalid)	[62]

[a] Omics repository databases.
[b] Annotation databases.

or transcriptome sequencing data. In a word, these computational tools will be helpful for the researchers to connect scattered pieces of evidence into meaningful hypotheses for further experimental validation [102]. Finally, by taking the *Dendrobium* genus plants as an example, we proposed a workflow for integrated analysis of the omics data. Since the phenotypic features were susceptible to the environmental cues, both genomic and chemical markers were recommended to be the indispensable criteria for plant authentication. Besides, in order to uncover the regulatory mechanisms underlying the biosynthesis pathways of dendrobine and polysaccharides, integrated analysis of genomics, transcriptomics, proteomics and metabolomics data was necessary. Notably, specific bioinformatics resources and toolkits were introduced into this research framework, showing their ability for efficient management of the omics data.

Taken together, we provided a comprehensive view of the currently available technologies and bioinformatics tools for the omics studies on medicinal plants. Multi-omics data-based inte-

Table 2. Non-exhaustive list of bioinformatics tools useful for medicinal plant research

Classification	Software name	Brief description	URL	Citation
Phenomics and taxonomics	TNRS	Taxonomic Name Resolution Service, an online application for automated and user-supervised standardization of plant scientific names	http://tnrs.iplantcollaborative.org/ (valid)	[63]
	ImageJ	Image processing and analysis platform written in Java	https://imagej.nih.gov/ij/ (valid)	[64]
	HTPheno	An image analysis pipeline for high-throughput plant phenotyping	http://htpheno.ipk-gatersleben.de/ (valid)	[65]
	PlantCV	Providing a collection of image analysis software for high-throughput plant phenotyping, which were integrated from a variety of source packages and algorithms	https://plantcv.danforthcenter.org/ (valid)	[66, 67]
Basic tools for genomics and/or transcriptomics studies	PLACE	Plant *cis*-acting regulatory DNA elements database	https://www.dna.affrc.go.jp/PLACE/?action=newplace (valid)	[68]
	WebLogo	Generating graphical representation of the sequence conservation pattern within a multiple sequence alignment	http://weblogo.berkeley.edu/logo.cgi (valid)	[69]
	PlantCARE	A database of plant cis-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences	http://bioinformatics.psb.ugent.be/webtools/plantcare/html/ (valid)	[70]
	Vienna RNA web server	Providing software packages for RNA secondary structure analyses, such as RNAfold	http://rna.tbi.univie.ac.at/ (valid)	[71]
	RNAshapes	A locally installed software for RNA secondary structure predictions	https://bibiserv.cebitec.uni-bielefeld.de/rnashapes (valid)	[72]
	Bowtie	An ultrafast, memory-efficient short read aligner	http://bowtie-bio.sourceforge.net/index.shtml (valid)	[73]
	Bowtie 2	An ultrafast and memory-efficient tool for aligning sequencing reads to reference sequences, especially for long read mapping to long reference sequences	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml (valid)	[74]
	Trinity	For *de novo* assembly of full-length transcripts based on RNA-seq data	https://github.com/trinityrnaseq/trinityrnaseq/wiki (valid)	[75]
	SOAPdenovo-Trans	Another tool for *de novo* transcriptome assembly from RNA-seq data	https://sourceforge.net/projects/soapdenovotrans/ (valid)	[76]
	TopHat + Cufflinks	For reference genome-based RNA-seq read alignment, full-length transcriptome assembly and comparative gene expression analysis	http://ccb.jhu.edu/software.shtml (valid)	[77]
	HISAT + StringTie + Ballgown	A newly improved software combination for reference genome-based RNA-seq read alignment, full-length transcriptome assembly and comparative gene expression analysis		[78]
	PEA	An integrated R toolkit for plant epitranscriptome analysis	https://hub.docker.com/r/malab/pea (valid)	[79]
	SAM	Significance analysis of microarray	http://www.biostat.umn.edu/&#x007E;baolin/research/ (valid)	[80]
ncRNA identification and design	miRPlant	An integrated tool for identification of plant miRNA from RNA sequencing data	http://sourceforge.net/projects/mirplant/ (valid)	[81]
	miRDeep-P	A computational tool for analyzing the miRNA transcriptome in plants	https://sourceforge.net/projects/mirdp/ (valid)	[82]
	PmiRDiscVali	An integrated pipeline for plant miRNA discovery and validation	https://github.com/unincrna/pmirdv (valid)	[83]
	P-SAMS	A web site for plant artificial miRNA and synthetic *trans*-acting small interfering RNA design	http://p-sams.carringtonlab.org/ (valid)	[84]
	NATpipe	An integrative pipeline for systematical discovery of natural antisense transcripts and phase-distributed nat-siRNAs from *de novo* assembled transcriptomes	www.bioinfolab.cn/NATpipe/NATpipe.zip (valid)	[85]
	PLncPRO	A bioinformatics tool for prediction of long ncRNAs in plants	http://ccbb.jnu.ac.in/plncpro/ (valid)	[86]
	PcircRNA_finder	A software for non-coding circular RNA prediction in plants	http://ibi.zju.edu.cn/bioinplant/tools/manual.htm (valid)	[87]

(Continued)

Table 2. Continued

Classification	Software name	Brief description	URL	Citation
Small RNA target prediction and validation	psRNATarget	A plant small RNA target analysis server	http://plantgrn.noble.org/psRNATarget/ (valid)	[88]
	TAPIR	A web server for the prediction of plant miRNA targets, including target mimics	http://bioinformatics.psb.ugent.be/webtools/tapir/ (valid)	[89]
	CleaveLand	A pipeline for degradome-seq data-based identification of cleaved small RNA targets	http://sites.psu.edu/axtell/software/cleaveland4/ (valid)	[90]
	sPARTA	A pipeline for degradome-seq data-based identification of cleaved miRNA targets	http://mpss.udel.edu/tools/mirna_apps/download.php (invalid)	[91]
	IIKmTA	Inter and intra kingdom miRNA-target analyzer	http://www.bioinformatics.org/iikmta/ (valid)	[92]
Functional analyses (gene annotations, networks and pathways)	PlantTFcat	An online plant transcription factor and transcriptional regulator categorization and analysis tool	http://plantgrn.noble.org/PlantTFcat/ (valid)	[93]
	Cytoscape	A software environment for integrated models of biomolecular interaction networks	http://www.cytoscape.org/ (valid)	[94]
	agriGO v2.0	An online server for GO term enrichment analysis	http://systemsbiology.cau.edu.cn/agriGOv2/ (valid)	[95]
	WGCNA	An R package for weighted correlation network analysis	https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/index.html (valid)	[96]
	INfORM	An R application enabling non-expert users to detect, evaluate and select gene modules with high statistical and biological significance	https://github.com/Greco-Lab/INfORM (valid)	[97]
	PathPred	An enzyme-catalyzed metabolic pathway prediction server	https://www.genome.jp/tools/pathpred/ (valid)	[98]
Integrated analyses	mixOmics	An R package offering a wide range of multivariate methods for integrated analysis of multiple types of omics data	http://mixomics.org/ (valid)	[99]
	iArray	Integrative Array Analyzer, a software package for analysis of cross-platform and cross-species microarray data	http://zhoulab.usc.edu/ (invalid)	[100]
	BRB-ArrayTools	An integrated R package for the visualization and statistical analysis of microarray gene expression, copy number, methylation and RNA-Seq data	https://brb.nci.nih.gov/BRB-ArrayTools/index.html (valid)	[101]

grated analyses should be an efficient approach to authenticate the medicinal plant materials and to decipher the secondary metabolic pathways in the medicinal plants.

Phenomics and taxonomics studies

Phenotyping, classification and authentication

The plant phenotypes are the outputs of gene–environment interactions. Under a particular growth condition, the phenotypes include all of the morphological and the physiological traits expressed from a plant genome [103]. Recording the phenotypic parameters from the single-cell level to the whole-plant level, termed 'phenotyping', is especially important for identification and classification of novel plant species and is also essential for mechanistic investigation of the genomic and the environmental effects on the plant phenotypes [104]. The vital role of phenotyping in plant biology leads to the birth of a sub-discipline termed 'phenomics', which includes controlled greenhouse or field experimentation, application of imaging technologies and development of the image analysis tools [103]. Generally, a plant phenomics study could be implemented by the following steps: (1) experimental design, such as setting parameters of water and nutrition supply, temperature, light intensity and humidity for plant growth, and design for biotic or abiotic stress treatment; (2) image acquisition by using a variety of cameras capturing signals from visible, ultra-violet and infrared spectra [105]; (3) image data management and interpretation, including data parameterization and storage, quantitative measurement of plant morphology (e.g. geometric and color properties), growth dynamics (e.g. seed germination, root elongation and plant weight) and physiological status (e.g. stress response, chlorophyll content, photosystem II quantum efficiency), and modeling. During the initial stage, both image acquisition and data interpretation heavily relied on manual labor, leading to unsatisfactory efficiency of phenotyping. With the technical advance, high-resolution imaging platforms and computer-assisted high-throughput analytical tools have been made available for the plant biologists, which have been introduced by several excellent reviews [105–107]. Some of the imaging platforms and bioinformatics tools were developed for specific purpose, such as root phenotyping [104, 108], stress phenotyping [109] and cellular functional analyses [110]. Notably, data from plant phenomics studies is not only valuable for model plant investigation and crop improvement [104, 111] but also precious for medicinal plant research. First, extraction of morphological features is indispensable for discrimination and classification of various medicinal plant species. Second, decoding the optimum cultivation conditions for medicinal plants should be one of the prerequisites for the mechanistic study on the formation of their authentic quality. Third, monitoring the key traits (e.g. the growing status of the medicinal parts) and recording the correlated parameters (e.g. the accumulation of the secondary metabolites) will pave a way for quality control (termed 'authentication') of the medicinal plants. Below, we will introduce several databases (Table 1) and analytical toolkits (Table 2) that are potentially useful for phenomics and taxonomics studies on the medicinal plants.

The newly updated database, PlantCLEF (2019), collected thousands of plant images taken from diverse visual angles, scales and organs [23]. In order to imitate a real-life automated system for plant identification and classification, the raw images were used by PlantCLEF. The convolutional neural network-based algorithm was adopted for machine learning, with the purpose of extracting common traits belonging to a defined plant genus or family. Currently, the PlantCLEF project only contains the images of 1000 plant species representative for the flora of Western Europe. However, it is conceivable that community-based collection of plant images will be initiated around the world, which might greatly facilitate the identification and classification of the novel medicinal plants. A compensatory database MPID (medicinal plant images database; https://library.hkbu.edu.hk/electronic/libdbs/mpd/) hosted by HongKong Baptist University is publicly available for phenomics studies on the medicinal plants. To our knowledge, MPID is the most comprehensive repository accommodating phenotypic data specifically for the medicinal plants at present. In addition to the representative images, the database also provides users with taxonomic names, morphological and environmental parameters and medicinal values of more than 1000 medicinal plant species. MPDB (medicinal plant database of Bangladesh) 1.0 is a more specific database for the medicinal plants distributed around Bangladesh [22]. The database accommodates 406 medicinal plants with their scientific and local names and medicinal parts for standardized nomenclature and taxonomy.

In addition to the databases storing phenotypic data of medicinal plants, toolkits for computer-assisted image processing and analysis have been reported. PlantCV (plant computer vision), a platform written in Python, provides a collection of the open-source and community-developed software packages and algorithms for high-throughput plant phenotyping [66, 67]. The integrated platform contains tools capable of analyzing images containing multiple plants, analyzing leaf segmentation by distance-based watershed transformation and analyzing plant shapes based on landmark identification. PlantCV is freely achievable from GitHub (https://github.com/danforthcenter/plantcv) [112]. ImageJ is another image processing and analysis platform. Different from PlantCV, ImageJ is written in Java [64]. Currently, the ImageJ platform is compatible for more than 500 plugins with diversified utilities, such as background correction, graphic segmentation, and normalization. HTPheno, implemented as one of the plugins for ImageJ, is a high-throughput color image analysis pipeline for determination of plant growth status and fitness [65]. For each plant, pictures taken from two different angles (i.e. top and side views) are subject to combinatorial analysis, in order to obtain the phenotypic parameters (e.g. height, width and projected shoot area) of the growing plants. The HTPheno pipeline is constituted of six major steps, including image retrieval, target region definition, plant segmentation, plant extraction, morphological construction and result outputs.

In addition to an ideal phenotyping system, a standardized nomenclature system is also crucial to improve the quality of the taxonomics studies. Misspelled species names and barbarous nomenclature will become an insurmountable obstacle to collect and integrate disparate data from ethnic pharmacopeias of different countries and will result in mismatched records and inflated numbers of medicinal plant species. TNRS (taxonomic name resolution service) is an online service for automated standardization of plant scientific names [63]. TNRS offers four optional taxonomic sources for users to perform taxonomic name standardization, including Tropicos (http://www.tropicos.org), Global Compositae Checklist (http://compositae.landcareresearch.co.nz/), USDA Plants (http://plants.usda.gov/java/) and NCBI Taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/). Compared to the other related applications, TNRS was claimed to have several advanced features (e.g. enabling batch processing and employing fuzzy matching algorithm), which endowed TNRS with higher efficiency of spelling error correction and spurious name elimination.

The bioinformatics resources specific for phenomics and taxonomics studies on the medicinal plants are still limited. Establishment of standardized pipeline for phenotypic data recording, construction of comprehensive databases for data storage and development of software packages for high-throughput data parsing should be essential for eliminating the bottleneck of systematic phenotyping and precise classification of the medicinal plants.

Genomics studies

Genome sequencing

As mentioned above, the hereditary genomic information is one of the major determinants for the morphological, physiological and biochemical outputs of a plant. Thus, uncovering the genetic codes by genome sequencing becomes a pivotal step toward in-depth studies on medicinal plants. To date, many research efforts have been devoted to the genome sequencing projects. For example, the draft genomes have been reported for *Rhodiola crenulata* [7], *Panax ginseng* [113], *Macleaya cordata* [8], *Glycyrrhiza uralensis* [9], *Chrysanthemum nankingense* [10], *Andrographis paniculata* [11], *Momordica charantia* [12] and *P. notoginseng* [14]. Besides, the complete chloroplast genomes of the *Glycyrrhiza* genus [114], *Artemisia annua* [115], *Forsythia suspensa* [116], *Amomum compactum* [117], *Saussurea involucrata* [118], *Salvia miltiorrhiza* [13], *Lycium chinense* Mill [119] and *Pogostemon cablin* [120] have also been released. Several genome databases of the medicinal plants have been made publicly available (Table 1). MPGR (medicinal plants genomics resource; http://medicinalplantgenomics.msu.edu/) accommodates the genomic information of 14 medicinal plant species, and HMOD (herbal medicine omics database) collects 22 herbal plant genomes for public download [46]. Both databases offer basic analytical tools to the users, such as 'Genome Browser' and 'BLAST'. Besides, several species-specific genome databases have also been constructed, such as MGH (Maca Genome Hub) [28] and HopBase [25]. Notably, in addition to the basic analytical tools for sequence alignment and gene identification, users could retrieve valuable resources from MGH for functional genomics studies. Specifically, the users could obtain the information about gene families, gene expression levels, ncRNAs and microRNA (miRNA)-mediated regulation from MGH [28].

DNA markers for authentication

In many cases, accurate identification and classification of a plant species not only can exclusively rely on its phenotypic characteristics but also should be performed at the molecular level [121]. It is especially important for authentication of medicinal plants, since authentic materials for drug preparation are the prerequisite for maintaining the consumers' belief and promoting healthy development of traditional medicine industry. One of the widely used molecular approaches for taxonomics studies is DNA barcoding, a DNA fingerprinting technique by using standardized genomic regions [122]. To date, several DNA barcodes, such as *matK*, *rbcL*, *trnH-psbA*, *ITS*, *trnL-F*, *5S-rRNA* and *18S-rRNA*, have been successfully applied for herbal plant identification and authentication. Accordingly, the DNA barcode databases such as MMDBD (medicinal materials DNA barcode database) [26] and its updated version [27] have been established (Table 1). In addition to the DNA barcodes, some other types of DNA markers, such as amplified fragment length polymorphisms, microsatellites, single nucleotide polymorphisms and random amplified polymorphic DNA, have also been applied for medicinal plant authentication [123]. Some specific databases have been established, such as CmMDb (*Cucumis melo* L. microsatellite database) [24]. With the release of new genomic data, it is conceivable that huge amounts of DNA markers will be developed for increasing numbers of medicinal plants. However, for efficient and accurate authentication, the DNA markers are recommended to be used in conjunction with other omics data such as metabolomics data [122], which will be introduced below.

Functional genomics

The values of genomic resources are not restricted to DNA marker-based taxonomics studies. The genomes with annotated genes provide the basic data for functional studies especially on secondary metabolic pathways related to the production of the bioactive ingredients in the medicinal plants. *Macleaya cordata* is a medicinal plant for the production of benzylisoquinoline alkaloids (e.g. sanguinarine and chelerythrine) with antimicrobial activities. Liu *et al.* [8] reported the 378 Mb draft genome of *M. cordata*, which contains 22 328 predicted protein-coding genes. Among these genes, 16 metabolic genes were discovered to be functionally involved in sanguinarine and chelerythrine biosynthesis, providing a knowledge foundation for bioengineering research on the metabolic pathways of benzylisoquinoline alkaloids. Neoandrographolide, one kind of diterpenoids with anti-inflammatory and anti-viral activity, is highly enriched in *A. paniculata*. By integrating Illumina short-read sequencing, PacBio long-read sequencing and Hi-C (high-confidence) sequencing platforms, Chen's group reported the 269 Mb draft genome assembly of *A. paniculata* [11]. A total of 25 428 protein-coding genes were annotated. Based on the integrated analysis of the genome and transcriptome sequencing data, the genes encoding diterpenoid synthases, cytochrome P450 monooxygenases, 2-oxoglutarate-dependent dioxygenases and UDP-dependent glycosyltransferases were identified to be potentially involved in the diterpenoid lactone biosynthesis pathway. In addition to the large-scale functional genomics studies introduced above, the fine-scale transgenic experiments including gene overexpression and targeted mutagenesis have also been performed in certain medicinal plants such as *S. miltiorrhiza* [18, 124].

There are several basic tools for the genome-based bioinformatics analyses (Table 2). Specifically, both PLACE [68] and PlantCARE [70] are the web servers capable of scanning for the potential *cis*-elements within the promoter regions of the query genes. WebLogo is a web-based application for the conservation analysis of amino acid or nucleic acid sequences. The consensus detected by a multi-sequence alignment is presented by a comprehensible graphical sequence logo [69]. It is especially useful for discovery of the potentially conserved motifs, such as the protein binding sites within a genomic region. Based on a manually curated list of the transcriptional regulators (TRs) in plants, Dai and his colleagues [93] developed a web-based tool PlantTFcat, which was particularly useful for genome-wide identification and categorization of TRs with high coverage and accuracy. Taken together, the above tools could be applied for the genome-based investigation of the transcriptional regulatory mechanisms in the medicinal plants.

Transcriptomics studies

Transcriptome-wide profiling

Profiling the spatio-temporal expression patterns of the plant genes could provide researchers with valuable hints for functional studies. In the early stage, expressed sequence tag (EST) data were frequently used for transcriptome-wide studies on the medicinal plants [125]. However, the limited coverage of EST sequencing data sets became an obstacle to tracking the weakly expressed genes. Later, probe hybridization-based microarray technology provides a much higher coverage for gene discovery and expression analysis [126]. The recent advent of the NGS technology enabled the researchers to investigate the transcriptional activities of the medicinal plant genes with unprecedented throughput and depth. Based on the NGS platform, several sequencing projects were carried out to examine the tissue- or organ-specific expression patterns of the herbal genes [15, 127–129]. Additionally, transcriptional dynamics at different developmental stages or treatment points has also been investigated. For example, Wang *et al.* [16] performed a transcriptome-wide profiling for the 5-year-old, 12-year-old and 18-year-old roots of *P. ginseng*, in order to obtain some valuable cues for the mechanistic study on ginsenoside biosynthesis. Gao and his colleagues [130] reported the leaf transcriptomes of *Ammopiptanthus mongolicus* under incremental drought stress duration.

To date, the transcriptome-wide studies have been performed for tens of the medicinal plants. The sequencing and microarray profiling data have been made publicly available in several databases, such as GEO (gene expression omnibus) [34], SRA (sequence read archive) [35], ArrayExpress [33], MPGR (http://medicinalplantgenomics.msu.edu/), HMOD [46], DsTRD (danshen transcriptional resource database) [29] and GarlicESTdb (garlic EST database) [30] (Table 1). Besides, there are several universal tools for basic research on plant transcriptomes (Table 2), which will be introduced below.

For microarray-based detection of differentially expressed genes, there are several basic packages, such as SAM (significance analysis of microarray) [80]. However, in many cases, an integrative meta-analysis is required for microarray-based studies. For example, independent array data sets are usually subject to horizontal meta-analysis in order to uncover the gene expression patterns under different biological conditions in the same plant species. In another case, microarray data should be analyzed in combination with phenomics, genomics, epigenomics and/or proteomics data. To this end, several toolkits have been made available for the microarray meta-analysis. iArray (Integrative Array Analyzer) is software package compatible for cross-platform and cross-species microarray data analyses [100]. Notably, iArray offers several functional modules for array data analyses, including data preprocessing, co-expression analysis, differential expression analysis, functional and transcriptional annotation and graphic visualization. BRB-ArrayTools is another integrated package for microarray data meta-analysis [101]. In addition to the microarray expression data, it also supports RNA-seq data, illumina methylation data and copy number data for meta-analysis. Similar to iArray, BRB-ArrayTools also provides user-friendly operating interface and generates graphical result outputs such as heat map, hierarchical clustering and KEGG (Kyoto encyclopedia of genes and genomes) pathways. For RNA-seq data analysis, there are two conditions. First, if the genomes are available for the medicinal plants, the combination of TopHat and Cufflinks [77] or the newly evolved combination of HISAT, StringTie and Ballgown [78] will be useful for RNA sequencing (RNA-seq) read alignment, full-length transcriptome assembly and comparative gene expression analysis. Second, in most cases, there is a lack of genomic information for the non-model plant studied. In this occasion, Trinity [75] or its substitutes such as SOAPdenovo-Trans [76] should be employed for *de novo* transcriptome assembly. To calculate the abundances of the assembled transcripts, RNA-seq read alignment should be performed by Bowtie [73] or Bowtie 2 [74], two efficient tools for sequence mapping. Notably, compared to Bowtie, Bowtie 2 is particularly suitable for aligning long sequencing reads (longer than 50 nt) to long reference sequences. Recently, the transcriptome-wide chemical modification profiling (termed epitranscriptomics) was considered to be another critical layer for mechanistic studies on gene regulation. The PEA (plant epitranscriptome analysis) toolkit developed by R language offers a series of efficient solutions for plant epitranscriptome analysis, including read mapping, modification signal calling, motif discovery and functional enrichment analysis of the gene set [79].

Co-expression, networks and functional analyses

The genes with similar expression patterns are often coordinately regulated at the transcriptional level, indicating that these co-expressed genes might be implicated in the functionally related pathways. For medicinal plants, it is particularly interesting to investigate the gene co-expression networks involved in secondary metabolism [131]. In the study by Sun *et al.* [132], co-expression analysis in *Lithospermum officinale* uncovered 20 unigenes encoding enzymes responsible for lithospermic/chlorogenic acid biosynthesis and 48 unigenes encoding enzymes for shikonin production. Based on the RNA-seq data, Chen's group constructed a weighted gene co-expression network constituted by 15 subnetworks in *Dioscorea nipponica*. Of these, 4 subnetworks consisting of 4665 genes were potentially involved in the regulation of dioscin biosynthesis [18].

To date, only a few databases related to the co-expression networks of the medicinal plants have been available. For example, in addition to the sequence information of the annotated gene families and the miRNAs, croFGD (*Catharanthus roseus* functional genomics database) also provides users with the information of the gene co-expression networks. Besides, croFGD offers the tools for the identification of the functional subnetworks [133]. AraNet was initially constructed to store the co-functional gene networks of *Arabidopsis thaliana*. Recently, it has been updated to AraNet v2, with improved genome coverage (~84% of the coding genome) and annotation accuracy. Notably, the co-functional gene networks of 28 non-model plant species such as *Glycine max* and *Vitis vinifera* were added to AraNet v2 [36]. By using an orthology-based projection of the non-model plant genes on the networks of *Arabidopsis*, it is hopeful that some of the medicinal plants will be included in AraNet. There are also several useful tools for network construction and functional analysis. Cytoscape is an open-source software environment with a core function for visual presentation of interactions or regulatory cascades, such as protein–protein interactions (PPIs) and genetic regulatory relationships [94]. The functionality of Cytoscape is extensible by integrating plug-ins with multiple utilities. For example, several key features such as expression profiles and chemical modifications could be integrated into the graphic networks. The online service agriGO is an easy-to-use tool for gene ontology (GO) term enrichment analysis of a plant gene set. Its updated version, agriGO v2.0, supports functional enrichment analysis for a total of 13 medicinal plant species [95]. Such functional analysis will facilitate the researchers to investigate the biological roles of a co-expressed gene set in the medicinal plants. Additionally, two R packages, WGCNA (weighted correlation network analysis) [96] and INfORM (inference of network response modules) [97] are useful for gene co-expression network analysis. The WGCNA package provides users with six major functions, including network construction, correlated module detection, functional gene selection, topological property calculation, data simulation and visualization. INfORM is also a comprehensive tool for identification of the biologically meaningful modules from certain networks. Different from WGCNA, the newly developed INfORM application presents an intuitive graphical interface for the non-expert users.

ncRNA discovery and functional studies

Another utility of transcriptome sequencing data is for discovery and functional analysis of the plant ncRNAs, such as the miRNAs and the long non-coding RNAs (lncRNAs). As one of the well-known small RNA species, the medicinal plant miRNAs have been demonstrated to possess pivotal regulatory roles in organ development and secondary metabolism [20, 134–137]. More intriguingly, several pieces of evidence indicated that the plant-originated miRNAs could be transmitted to the mammalian system and perform cross-kingdom regulation of specific targets [138]. Indeed, several reports have shown the therapeutic values of plant miRNAs in treating cancers [139–141] and suppressing influenza A virus infection [142]. On the other hand, the lncRNAs have been identified by transcriptome-wide studies in *S. miltiorrhiza*, *Digitalis nervosa*, *Ginkgo biloba* and *P. ginseng* [21, 143–145]. Some of these lncRNAs were reported to be involved in stress response [19, 144] or small interfering RNA (siRNA) production [145]. Interestingly, some of the lncRNAs were found to be the downstream targets of specific miRNAs [21], making gene regulatory network to be much more complicated than previously thought.

Recently, our group has established a miRNA database specific for medicinal plants. The current version of MepmiRDB (medicinal plant microRNA database) [32] provides sequence, expression and target information of the miRNAs identified from 29 medicinal plant species. Notably, the construction of MepmiRDB relied on several bioinformatics tools. First, based on the mature miRNAs of Viridiplantae registered in miRBase [31] and the small RNA sequencing (sRNA-seq) data of the medicinal plants, sequence identity-based search was performed to identify the conserved miRNA candidates in the medicinal plants. Second, based on the transcriptome assembly and the sRNA-seq data, novel miRNAs, especially for the species-specific ones, could be predicted by using the computational tools such as miRPlant [81] and miRDeep-P [82]. Then, both RNAshapes [72] and RNAfold [71] could be used to check the secondary structures of the predicted miRNA precursors, since most of the precursors could form hairpin-like structures. Different from the locally installed RNAshapes, RNAfold is an online service for secondary structure prediction. Thus, it is independent on the local machine memory and is capable of treating the query sequences less than 10sRNA-seq000 nt. PmiRDiscVali [83] is another useful software package for plant miRNA identification. Different from miRPlant and miRDeep-P, the secondary structures and the sRNA distribution patterns of the miRNA precursors could be graphically presented by the outputs of PmiRDiscVali. Besides, based on degradome sequencing (degradome-seq) data, the detected processing signals will be marked at the ends of the mature miRNA-coding regions on the precursors. Together, the three lines of visible evidence provided by PmiRDiscVali could facilitate the users to make a judgment on the reliability of the miRNA candidates. Third, miRNA target prediction could be performed by using psRNATarget [88], a popular tool for plant sRNA target prediction. TAPIR [89] is another alternative online service for *in silico* identification of the miRNA targets. Different from psRNATarget, TAPIR could also be used to predict target mimics. As mentioned above, some of the plant endogenous miRNAs are implicated in cross-kingdom regulation. In this regard, IIKmTA [92], developed for both inter- and intra-kingdom identification of miRNA–target pairs, could be useful for the discovery of the herbal miRNA targets in the mammalian system. Finally, based on degradome-seq data, the predicted miRNA–target pairs could be validated by using CleaveLand [90] or sPARTA [91]. As a result, the validated miRNA–target interactions form a regulatory network that could be drawn by using Cytoscape [94].

In addition to the computational tools for miRNA studies, some other software packages might be valuable for the identification and functional analyses of the other ncRNA species. For example, based on the transcriptome assembly and the sRNA-seq data, NATpipe [85] could be employed for large-scale identification of nature antisense transcripts (NATs) and the associated siRNAs. Artificial miRNAs (amiRNAs) and synthetic *trans*-acting siRNAs (syn-tasiRNAs) are efficient transgenic tools for targeted gene repression in plants. P-SAMS (plant small RNA maker site) is a web tool providing two applications, 'amiRNA Designer' and 'syn-tasiRNA Designer', for the simplified and automated design of amiRNAs and syn-tasiRNAs, respectively [84]. To our knowledge, no database related to the lncRNAs of medicinal plants has been publicly available. However, there are several bioinformatics tools that could be used for ncRNA identification, such as PLncPRO for lncRNA discovery [86] and PcircRNA_finder for circular RNA (circRNA) prediction [87].

Proteomics studies

Qualitative and quantitative analyses

The functions of most protein-coding genes are implemented by their protein products. In this regard, identification and quantitative analysis of the gene products by proteome profiling is especially important for the researchers to gain deep insights into the mechanisms underlying developmental and metabolic processes of the medicinal plants. In Jacobs *et al.*'s study [146], 2D polyacrylamide gel electrophoresis was performed for systematic analysis of the proteome of *C. roseus*. Facilitated by mass spectrometry, several proteins participated in alkaloid biosynthesis were identified, such as strictosidine synthase and tryptophan synthase [146]. The seed germination of *Dendrobium officinale* is stimulated by the colonization of mycorrhizal fungi. Based on integrated analysis of the transcriptomic and the proteomic data, Chen and his colleagues [147] examined the molecular changes during symbiotic germination of the orchid seeds. The biological functions of a protein not only are determined by its linear amino acid sequence but also rely on its structure. Besides, the stability and the activity of a protein could be influenced by post-translational modifications, such as phosphorylation, acetylation, crotonylation and succinylation. To date, several studies have been carried out for protein structure interpretation [148–150] and protein modification profiling [151, 152], which provided the mechanistic insights into the developmental or the metabolic processes of the medicinal plants.

To our knowledge, there are few proteomics databases specific for the medicinal plants. However, the worldwide protein knowledgebases, such as UniProt [37] and Pfam [38], might be useful for the identification and functional annotation of the medicinal plant proteins. PPIs contribute to a central module of the complex biological network. Besides, the unknown functions of a novel protein might be partially inferred from its interaction partners [153]. In this regard, it is interesting to investigate the potential interactions among the proteins identified in medicinal plants. For this purpose, orthology-based projection of the non-model plant proteins on the PPIs of the model plants will enable the researchers to construct the PPI networks in the medicinal plants. To this end, the databases storing PPI information of the model plants, such as BioGRID (biological general repository for interaction datasets) [39], STRING (search tool for the retrieval of interacting genes) [40], IntAct [41], PAIR (predicted *Arabidopsis* interactome resource) [42] and PRIN (predicted rice interactome network) [43], could be useful.

Metabolomics studies

Metabolome profiling and metabolite-based authentication

Compared to animals and microorganisms, plants can produce much more bioactive metabolites. Various plant metabolites, such as alkaloids, anthocyanins, flavonoids, quinines and terpenoids, have been verified to possess immense medicinal values [154]. To date, several analytical technologies, including gas chromatography–mass spectrometer, liquid chromatography–mass spectrometer, Fourier transform–infrared spectrometer, nuclear magnetic resonance, capillary electrophoresis–mass spectrometry and liquid chromatography–photodiode array, have been applied for identification, annotation and quantitative profiling of the plant metabolites (see review in [154]). In addition to the identification of medicinal constituents, the metabolomics studies are also important for quality assessment of the medicinal plant materials [155, 156]. For example, by using UHPLC-TOFMS (ultra-high performance liquid chromatography combined with time-of-flight mass spectrometry), an untargeted metabolomic method, Afzan and his colleagues [157] identified 15 glycosylated flavone markers that were capable of distinguishing different varieties of *Ficus deltoidea*. Interestingly, the results of the chemotaxonomy-based differentiation were highly consistent with those of the genome-based taxonomics studies on *F. deltoidea*. Moreover, for a given variety of *F. deltoidea*, the stability of the chemical markers was not influenced by geographical locations and growing conditions of the plant [157]. From this point of view, the metabolic markers are one of the competent tools for medicinal plant authentication.

Bioinformatics-assisted studies on metabolic pathways and medicinal ingredients

Understanding of the regulatory cascades related to plant metabolism is the first step toward genetic improvement and metabolic engineering of the medicinal plants. In this section, we will introduce the databases and the software that are useful for the researchers to draw a specific metabolic pathway. Besides, the bioinformatics resources valuable for plant-based drug discovery will also be described.

As introduced in the previous section, MPGR (http://medicinalplantgenomics.msu.edu/) provides users with the genomic information of 14 medicinal plant species. Notably, it also offers rich information related to the bioactive constituents discovered in these plants, including the chemical structures of these constituents, the key enzymes and the coding genes for the biosynthesis of these constituents and the related references. HMOD is a comprehensive database accommodating multiple types of the omics data from diverse medicinal herbs [46]. It contains two functional modules related to metabolomics studies: (1) The 'Pathways' module provides information about the biosynthetic pathways and the catalytic enzymes related to the medicinal ingredients (e.g. glucosinolate, betalain, caffeine and isoflavonoid), which is linked to the KEGG PATHWAY Database [45]. (2) The 'Metabolomics' module collected and summarized the published metabolomics studies on the medicinal plants, including institutes, materials and methods and research conclusions. *Catharanthus roseus* was reported to synthesize terpenoid indole alkaloids (TIAs) with pharmaceutical importance. A specific database, CathaCyc, has been established to provide users with 390 pathways involving 1347 enzymes related to the synthesis of TIAs and their metabolic precursors in *C. roseus* [44]. Besides, the expression patterns of the enzyme-coding genes or other key regulators involved in the metabolism of *C. roseus* were derived from the analysis of the RNA-seq data, which could be visualized through bar diagrams. PathPred is a powerful web server that can be used to predict the multi-step synthetic pathway of a given query compound [98]. Based on the hypothesis that the structurally related compounds might share a generalized pathway, the prediction of the chemical reactions of a given compound is complemented through a similarity search against the KEGG COMPOUND database [158]. However, this similarity-based approach may lead to the failure in deciphering specialized pathways of certain derivatives.

One of the ultimate goals of metabolomics studies is linking a specific plant chemical constituent to its therapeutic value. However, many traditional prescriptions are prepared from a mixture of crude plant materials, resulting in a quite complicated scene that a suite of metabolites with distinct chemical properties match the same pharmacological target. In this regard, construction of the public databases reporting the bioactive compounds from the medicinal plants should be the first step toward the precision of the traditional medicine industry. To date, several repositories accommodating the chemical structures and/or the therapeutic values of the bioactive ingredients identified from the medicinal plants have been publicly available (Table 1), such as SACPD (Saudi anti-human cancer plants database) [58], MAPS (medicinal plant activities, phytochemicals & structural database) [53], DIACAN (the antidiabetic and anticancer medicinal plants database) [57], AromaDb [47], NeMedPlant [54], IMPPAT [51], Phytochemica [48], SerpentinaDB [49], InDiaMed [52], KNApSAcK [60] and HIT (herb ingredients' targets) [62]. For traditional Chinese medicine (TCM) research, TCM Database@Taiwan provides both 2D and 3D structure information of more than 20 000 pure compounds isolated from the medicinal plants [50]. In order to facilitate the researchers to deduce the regulatory effects of TCMs on gene expression, TCMGeneDIT collected the information on diseases, TCM-target genes, target gene associated signaling pathways, and PPIs through text mining [61]. Relying on the increasing availability of the metabolomics data, more comprehensive databases not restricted to TCMs have been constructed during the past years. The CMAUP (collective molecular activities of useful plants) database provides the information of 47 645 ingredients identified from 5645 plants, including 646 therapeutic targets assigned to 234 KEGG pathways [56]. NPASS (natural product activity and species source) is another database storing 35 032 natural products (NPs) associated with 5863 targets (including 2946 proteins, 1352 microbial species and 1227 cell lines) [55]. Notably, for the effects of the NPs on the biological targets, NPASS contains a total of 446 552 quantitative activity records for the NP–target pairs. TarNet is a database gathering the information on NP–protein interactions and PPIs related to specific biological pathways [59]. Notably, by mapping the query genes or proteins to the PPI database, TarNet can facilitate researchers to construct the networks related to specific diseases.

A workflow for integrated omics studies: a case study on *Dendrobium*

In the above sections, the recent advances of the omics studies on medicinal plants were introduced separately, including phenomics, taxonomics, genomics, transcriptomics, proteomics and metabolomics (Figure 1). However, in many cases, a research task, such as plant authentication [159] and mechanistic studies on plant metabolism [160], should be implemented by integrated analysis of multiple types of the omics data. To date, several omics studies have been performed for the *Dendrobium* genus plants with medicinal and ornamental values [161]. In this case study, by taking *Dendrobium* as an example, we will propose an bioinformatics-assisted workflow of integrated omics studies on the medicinal plants.

As shown in Figure 2A, three *Dendrobium* species including *D. officinale* (Dof), *D. huoshanense* (Dhu) and *D. williamsonii* (Dwi) have distinguishable phenotypes such as plant height and leaf shape. However, it becomes difficult to distinguish tens of the *Dendrobium* species distributed in the nature. In this regard, both the morphological and the physiological data of each *Dendrobium* species should be recorded and stored in the public databases such as MPID (https://library.hkbu.edu.hk/electronic/libdbs/mpd/). These data sets should be useful for taxonomics studies and plant authentication. However, some of the phenotypic features might be susceptible to environmental cues. Besides, only the dried stems of certain *Dendrobium* species are

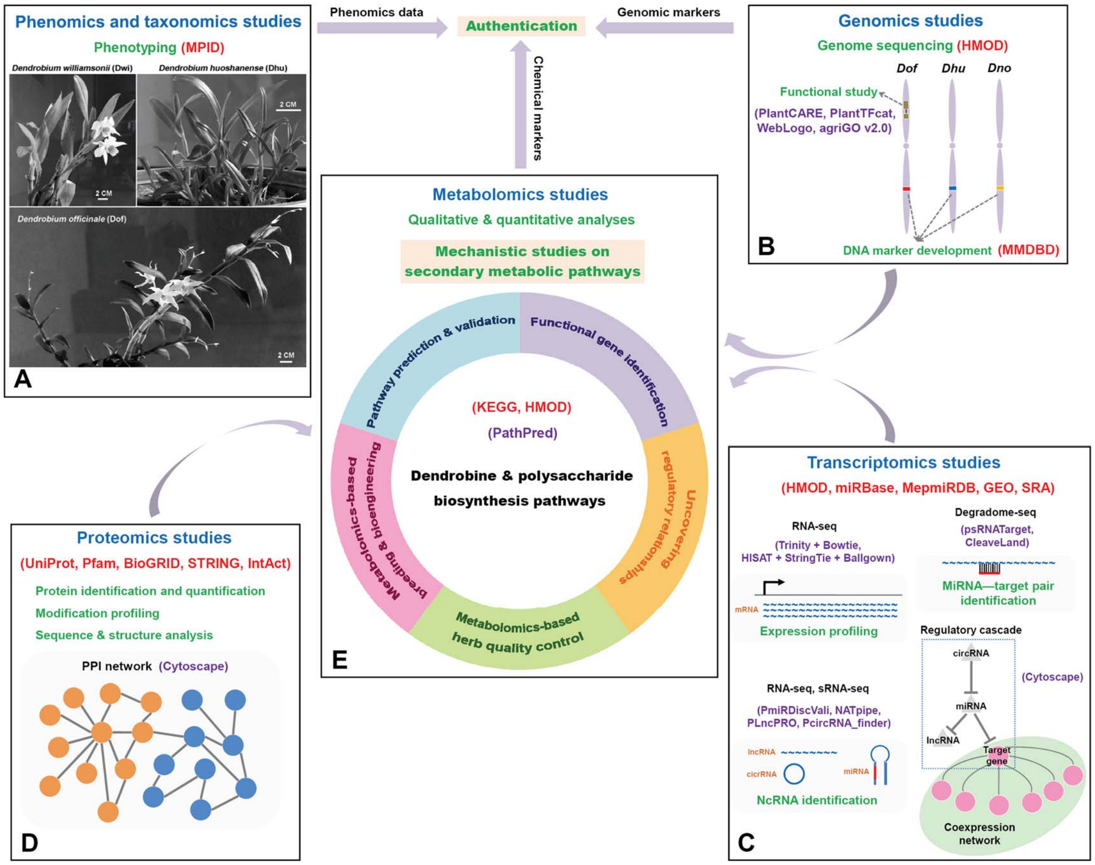
Phenomics data
Authentication
Genomic markers
Chemical markers
Phenomics and taxonomics studies
Phenotyping (MPID)
Dendrobium williamsonii (Dwi)
Dendrobium huoshanense (Dhu)
2 CM
2 CM
Dendrobium officinale (Dof)
2 CM
A
Genomics studies
Genome sequencing (HMOD)
Dof
Dhu
Dno
Functional study
(PlantCARE, PlantTFcat, WebLogo, agriGO v2.0)
B
DNA marker development (MMDBD)
Metabolomics studies
Qualitative & quantitative analyses
Mechanistic studies on secondary metabolic pathways
Pathway prediction & validation
Functional gene identification
Uncovering regulatory relationships
Metabolomics-based herb quality control
Metabolomics-based breeding & bioengineering
(KEGG, HMOD)
(PathPred)
Dendrobine & polysaccharide biosynthesis pathways
E
Proteomics studies
(UniProt, Pfam, BioGRID, STRING, IntAct)
Protein identification and quantification
Modification profiling
Sequence & structure analysis
PPI network (Cytoscape)
D
Transcriptomics studies
(HMOD, miRBase, MepmiRDB, GEO, SRA)
RNA-seq
(Trinity + Bowtie, HISAT + StringTie + Ballgown)
mRNA
Expression profiling
Degradome-seq
(psRNATarget, CleaveLand)
MiRNA—target pair identification
Regulatory cascade
circRNA
miRNA
lncRNA
Target gene
(Cytoscape)
RNA-seq, sRNA-seq
(PmiRDiscVali, NATpipe, PLncPRO, PcircRNA_finder)
lncRNA
circRNA
miRNA
NcRNA identification
Coexpression network
C

Figure 2. The research workflow for the multi-omics data-based integrated studies on the *Dendrobium* genus. Three species belonging to the *Dendrobium* genus, i.e. *D. officinale* (Dof), *D. huoshanense* (Dhu) and *D. williamsonii* (Dwi), were included in the case study. (A) Phenomics and taxonomics studies. According to the photographs, the three *Dendrobium* species possess distinguishable morphological features, such as plant height and leaf shape. (B) Genomics studies, including genome sequencing, DNA marker development and functional genomics analyses. (C) Transcriptomics studies, including gene expression profiling, ncRNA discovery, miRNA-target identification and network construction. (D) Proteomics studies, such as identification and quantification of the functional proteins, and construction of PPI networks. (E) Metabolomics studies, such as pathway prediction and validation, and metabolomics-based molecular breeding and bioengineering. According to the research framework, precise authentication of the *Dendrobium* species requires integrated analysis of phenomics, genomics and metabolomics data. On the other hand, functional genomics studies, transcriptional regulatory networks and PPIs will contribute greatly for deciphering the secondary metabolic pathways in *Dendrobium*. Notably, specific databases (highlighted in red color) and bioinformatics toolkits (highlighted in purple color) have been introduced into the workflow.

used as a crude drug in TCM. Both cases will lead to an obstacle for phenomics data-based plant material authentication. Fortunately, diverse types of DNA markers have been developed and proved to be a powerful tool for classification and authentication of the *Dendrobium* species [162–164] (Figure 2B). MMDBD, a DNA barcode database of medicinal materials, accommodates the information of DNA markers for a total of 39 *Dendrobium* species [26]. Besides, MMDBD is compatible for new data submission.

The draft genome of *D. officinale* (also named *D. catenatum*) has been release recently [165, 166] (Figure 2B). It is accessible through the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/Dendrobium_catenatum/) or through the database HMOD [46]. Notably, HMOD provides users with the 'GBrowse' tool to search for a specific genomic position or the annotated gene models. In recent years, fine-scale functional studies were carried out in *Dendrobium* through double-stranded RNA-mediated gene silencing [167], CRISPR/Cas9-based gene editing [168] or gene over-expression [169]. Relying on the genome availability of *D. officinale*, large-scale functional analyses could be implemented by using bioinformatics tools. For example, PlantCARE could be used for *cis*-element discovery within the promoter regions [70]. PlantTFcat could be used for identification and categorization of the TRs [93]. WebLogo could be applied for conserved sequence motif discovery [69]. Moreover, agriGO v2.0 could be employed for gene function enrichment analysis [95].

The medicinal and economic value of *Dendrobium* inspired more and more research efforts on the mechanistic study on gene regulation. Transcriptome-wide profiling enabled the researchers to obtain a global view of the spatio-temporal expression patterns of the *Dendrobium* genes [15]. To date, dozens of the transcriptome sequencing (including RNA-seq, sRNA-seq and degradome-seq) data sets of *Dendrobium* have been available in the public databases, such as HMOD [46], MepmiRDB [32], GEO [34] and SRA [35] (Figure 2C). Based on the genomic information of *D. officinale*, the software package 'HISAT + StringTie + Ballgown' could be used for reference-based transcriptome assembly and gene expression profiling [78]. However, it is not always the case. Recently, RNA-seq experiment was performed for the three *Dendrobium* species, including Dof, Dhu and Dwi (unpublished data). By mapping the RNA-seq reads onto the Dof genome, we discovered that the mapping ratio of Dwi (~25%) was much lower than those of Dof (~86%) and Dhu (~67%). The result indicates that, compared to Dwi, Dhu might be more closely related to Dof. From another angle, transcriptome assembly of Dwi might be failed by treating the Dof genome as the reference. At this time, a different software package 'Trinity [75] + Bowtie

[74]' should be used for *de novo* transcriptome assembly and investigation of the gene expression patterns.

Also based on the RNA-seq data of *Dendrobium*, both lncRNAs and circRNAs could be predicted by using PLncPRO [86] and PcircRNA_finder [87], respectively (Figure 2C). The availability of the *Dendrobium* sRNA-seq data allows the researchers to perform a transcriptome-wide search for the miRNAs and the NAT-derived siRNAs by using PmiRDiscVali [83] and NATpipe [85], respectively. Moreover, psRNATarget [88] could be utilized for sRNA target prediction, and CleaveLand [90] could be used for target validation based on the degradome-seq data. Finally, gene regulatory networks constituted by ncRNA–target pairs could be drawn by using Cytoscape [94] (Figure 2C).

For the proteomics studies on *Dendrobium*, the protein reference databases such as UniProt [37] and Pfam [38] could be used to annotate the *Dendrobium* protein-coding genes. Besides, BioGRID [39], STRING [40] and IntAct [41] are valuable to identify the PPIs from the *Dendrobium* proteomes (Figure 2D).

The medicinal value of *Dendrobium* was attributed to its secondary metabolites such as dendrobine and polysaccharides. Based on the data from the genomics, transcriptomics and proteomics studies, certain functional genes (e.g. enzyme-coding genes and key regulators) involved in the biosynthesis of dendrobine and polysaccharides might be inferred from the KEGG PATHWAY Database [45]. Besides, the related synthetic pathways might be predicted by using PathPred [98] (Figure 2E).

To date, several R packages have been available for the integrated analyses of multiple omics data types and have been summarized by a recent review [170]. Here, we would like to introduce an R package mixOmics [99], which is a useful pipeline for omics data integration and biological model prediction. Compared to the other packages, mixOmics have several superiorities. Firstly, it is compatible for both single and multiple omics data set analyses. sPLS-DA (sparse partial least square-discriminant analysis) could be used for supervised analysis of a single omics data set. DIABLO enables *N*-integration analysis of different types of omics data (such as transcriptomics, proteomics and metabolomics data) from certain biological samples. Differently, MINT enables *P* integration of omics data sets from several independent studies. Secondly, mixOmics provides novel sparse variants to the users, which enables feature selection. Thirdly, mixOmics offers several key functions, such as plotIndiv, plotArrow, network, cim and circosPlot, for users to present their selected features with insightful graphical outputs.

Summarily, centered by metabolomics studies, a research framework was proposed for *Dendrobium* by integrating different types of the omics data (Figure 2). We raised the opinion that (1) authentication of the plant materials from *Dendrobium* should be performed by integrated analysis of the phenomics, genomics and metabolomics data and (2) the multi-omics data-based mechanistic studies on secondary metabolism pave a way for molecular breeding- and/or bioengineering-based quality improvement of *Dendrobium*.

Concluding remarks and further perspectives

The wide application of the high-throughput technologies and the associated analytical tools has greatly accelerated the progress of the omics studies. In this paper, we summarized the recent achievements of the omics studies specifically for the medicinal plants. Facilitated by the public data repositories and the computational tools, multi-omics data-based integrated approaches were recommended for medicinal plant research, such as plant authentication, and mechanistic studies on plant metabolism. However, omics studies on the medicinal plants are still at the early stage. Further research efforts should be made from many aspects. First, increasing pieces of evidence pointed to the non-negligible influence of epigenetic modifications on gene expression. It will be interesting to perform epigenome and epitranscriptome sequencing to investigate the effects of epigenetic modifications on the secondary metabolic pathways of the medicinal plants. Second, discovery and functional analysis of the novel ncRNAs, such as lncRNAs and circRNAs, will contribute another regulatory layer to the gene regulatory networks in the medicinal plants. It is hopeful that the ncRNA databases specific for the medicinal plants, such as MepmiRDB [32], will be constructed in the near future. Third, as shown in Table 1, several previously published databases are no longer accessible. Considering their value for medicinal plant research, we encourage the corresponding research teams to reactivate these databases. Additionally, the currently available bioinformatics resources should be under routine maintenance and be continuously updated.

Funding

This work was funded by the National Natural Science Foundation of China (81730108, 81973635 and 31601062), Major Increase Or Decrease Program In The Central Finance Level (2060302) and Zhejiang Provincial Key Research and Development Project (grants 2017C02011 and 2018C02030).

Key Points

- The recent advances of omics studies on the medicinal plants were summarized from several aspects, including phenomics and taxonomics, genomics, transcriptomics, proteomics and metabolomics.
- Integrated analysis of the omics data is important for authentication and mechanistic studies on secondary metabolism of the medicinal plants.
- Computational tools for proper storage, efficient processing and high-throughput analysis of the omics data were introduced.
- A case study was performed to show the utility of the analytical workflow for integrated omics studies.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

References

1. Strohl WR. The role of natural products in a modern drug discovery program. *Drug Discov Today* 2000;**5**:39–41.
2. Buriani A, Garcia-Bermejo ML, Bosisio E, *et al*. Omic techniques in systems biology approaches to traditional Chinese medicine research: present and future. *J Ethnopharmacol* 2012;**140**:535–44.
3. Yuan Y, Lee H, Hu H, *et al*. Single-cell genomic analysis in plants. *Genes* 2018;**9**.
4. Fukushima A, Kusano M, Redestig H, *et al*. Integrated omics approaches in plant systems biology. *Curr Opin Chem Biol* 2009;**13**:532–8.

5. Pinu FR, Beale DJ, Paten AM, *et al*. Systems biology and multi-omics integration: viewpoints from the metabolomics research community. *Metabolites* 2019;**9**:76.
6. Bassel GW, Gaudinier A, Brady SM, *et al*. Systems analysis of plant functional, transcriptional, physical interaction and metabolic networks. *Plant Cell* 2012;**24**:3859–75.
7. Fu Y, Li L, Hao S, *et al*. Draft genome sequence of the Tibetan medicinal herb *Rhodiola crenulata*. *Gigascience* 2017;**6**:1–5.
8. Liu X, Liu Y, Huang P, *et al*. The genome of medicinal plant *Macleaya cordata* provides new insights into benzylisoquinoline alkaloids metabolism. *Mol Plant* 2017;**10**:975–89.
9. Mochida K, Sakurai T, Seki H, *et al*. Draft genome assembly and annotation of *Glycyrrhiza uralensis*, a medicinal legume. *Plant J* 2017;**89**:181–94.
10. Song C, Liu Y, Song A, *et al*. The *Chrysanthemum nankingense* genome provides insights into the evolution and diversification of *Chrysanthemum* flowers and medicinal traits. *Mol Plant* 2018;**11**:1482–91.
11. Zhang D, Jiang C, Huang C, *et al*. The light-induced transcription factor FtMYB116 promotes accumulation of rutin in *Fagopyrum tataricum*. *Plant Cell Environ* 2019;**42**:1340–51.
12. Urasaki N, Takagi H, Natsume S, *et al*. Draft genome sequence of bitter gourd (*Momordica charantia*), a vegetable and medicinal plant in tropical and subtropical regions. *DNA Res* 2017;**24**:51–8.
13. Xu H, Song J, Luo H, *et al*. Analysis of the genome sequence of the medicinal plant *Salvia miltiorrhiza*. *Mol Plant* 2016;**9**:949–52.
14. Zhang D, Li W, Xia EH, *et al*. The medicinal herb *Panax notoginseng* genome provides insights into ginsenoside biosynthesis and genome evolution. *Mol Plant* 2017;**10**:903–7.
15. Meng Y, Yu D, Xue J, *et al*. A transcriptome-wide, organ-specific regulatory map of *Dendrobium officinale*, an important traditional Chinese orchid herb. *Sci Rep* 2016;**6**:18864.
16. Wang K, Jiang S, Sun C, *et al*. The spatial and temporal transcriptomic landscapes of ginseng, *Panax ginseng* C, a Meyer. *Sci Rep* 2015;**5**:18283.
17. Rai A, Nakaya T, Shimizu Y, *et al*. De novo transcriptome assembly and characterization of *Lithospermum officinale* to discover putative genes involved in specialized metabolites biosynthesis. *Planta Med* 2018;**84**:920–34.
18. Sun W, Wang B, Yang J, *et al*. Weighted gene co-expression network analysis of the dioscin rich medicinal plant *Dioscorea nipponica*. *Front Plant Sci* 2017;**8**:789.
19. Li D, Shao F, Identification LS. Characterization of mRNA-like noncoding RNAs in *Salvia miltiorrhiza*. *Planta* 2015;**241**:1131–43.
20. Vashisht I, Mishra P, Pal T, *et al*. Mining NGS transcriptomes for miRNAs and dissecting their role in regulating growth, development, and secondary metabolites production in different organs of a medicinal herb *Picrorhiza kurroa*. *Planta* 2015;**241**:1255–68.
21. Wang L, Xia X, Jiang HR, *et al*. Genome-wide identification and characterization of novel lncRNAs in *Ginkgo biloba*. *Trees* 2018;**32**:1429–42.
22. Ashraf MA, Khatun A, Sharmin T, *et al*. MPDB 1.0: a medicinal plant database of Bangladesh. *Bioinformation* 2014;**10**:384–6.
23. Seeland M, Rzanny M, Boho D, *et al*. Image-based classification of plant genus and family for trained and untrained plant species. *BMC Bioinformatics* 2019;**20**:4.
24. Bhawna CPK, Bonthala VS, *et al*. CmMDb: a versatile database for *Cucumis melo* microsatellite markers and other horticulture crop research. *PLoS One* 2015;**10**:e0118630.
25. Hill ST, Sudarsanam R, Henning J, *et al*. HopBase: a unified resource for Humulus genomics. *Database (Oxford)* 2017;**2017**.
26. Lou SK, Wong KL, Li M, *et al*. An integrated web medicinal materials DNA database: MMDBD (medicinal materials DNA barcode database). *BMC Genomics* 2010;**11**:402.
27. But GW WTH, Wu HY, *et al*. Medicinal materials DNA barcode database (MMDBD) version 1.5-one-stop solution for storage, BLAST, alignment and primer design. *Database (Oxford)* 2018;**2018**.
28. Chen J, Zhang J, Lin M, *et al*. MGH: a genome hub for the medicinal plant maca (*Lepidium meyenii*). *Database (Oxford)* 2018;**2018**.
29. Shao Y, Wei J, Wu F, *et al*. DsTRD: Danshen transcriptional resource database. *PLoS One* 2016;**11**:e0149747.
30. Kim DW, Jung TS, Nam SH, *et al*. GarlicESTdb: an online database and mining tool for garlic EST sequences. *BMC Plant Biol* 2009;**9**:61.
31. Griffiths-Jones S, Grocock RJ, van Dongen S, *et al*. miRBase: microRNA sequences targets and gene nomenclature. *Nucleic Acids Res* 2006;**34**:D140–4.
32. Meng Y, Gou L, Chen D, *et al*. PmiRKB: a plant microRNA knowledge base. *Nucleic Acids Res* 2011;**39**:D181–7.
33. Brazma A, Parkinson H, Sarkans U, *et al*. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 2003;**31**:68–71.
34. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;**30**:207–10.
35. Kodama Y, Shumway M, Leinonen R, *et al*. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res* 2012;**40**:D54–6.
36. Lee T, Yang S, Kim E, *et al*. AraNet v2: an improved database of co-functional gene networks for the study of *Arabidopsis thaliana* and 27 other nonmodel plant species. *Nucleic Acids Res* 2015;**43**:D996–1002.
37. UniProt C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**:D506–15.
38. El-Gebali S, Mistry J, Bateman A, *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res* 2019;**47**: D427–32.
39. Oughtred R, Stark C, Breitkreutz BJ, *et al*. The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 2019;**47**:D529–41.
40. Szklarczyk D, Franceschini A, Wyder S, *et al*. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015;**43**:D447–52.
41. Kerrien S, Aranda B, Breuza L, *et al*. The IntAct molecular interaction database in 2012. *Nucleic Acids Res* 2012;**40**:D841–6.
42. Lin M, Shen X, Chen XPAIR. The predicted Arabidopsis interactome resource. *Nucleic Acids Res* 2011;**39**:D1134–40.
43. Gu H, Zhu P, Jiao Y, *et al*. PRIN: a predicted rice interactome network. *BMC Bioinformatics* 2011;**12**:161.
44. Van Moerkercke A, Fabris M, Pollier J, *et al*. CathaCyc, a metabolic pathway database built from *Catharanthus roseus* RNA-Seq data. *Plant Cell Physiol* 2013;**54**:673–85.
45. Kanehisa M, Furumichi M, Tanabe M, *et al*. KEGG: new perspectives on genomes, pathways diseases and drugs. *Nucleic Acids Res* 2017;**45**:D353–61.
46. Wang X, Zhang J, He S, *et al*. HMOD: an omics database for herbal medicine plants. *Mol Plant* 2018.
47. Kumar Y, Prakash O, Tripathi H, *et al*. AromaDb: a database of medicinal and aromatic plant's aroma molecules with

phytochemistry and therapeutic potentials. *Front Plant Sci* 2018;**9**:1081.

48. Pathania S, Ramakrishnan SM, Bagler G. Phytochemica: a platform to explore phytochemicals of medicinal plants. *Database* 2015;**2015**.
49. Pathania S, Ramakrishnan SM, Randhawa V, *et al.* SerpentinaDB: a database of plant-derived molecules of *Rauvolfia serpentina*. *BMC Complement Altern Med* 2015;**15**:262.
50. Chen CYTCM. Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. *PLoS One* 2011;**6**:e15939.
51. Mohanraj K, Karthikeyan BS, Vivek-Ananth RP, *et al.* IMPPAT: a curated database of Indian medicinal plants, phytochemistry and therapeutics. *Sci Rep* 2018;**8**:4329.
52. Tota K, Rayabarapu N, Moosa S, *et al.* InDiaMed: a comprehensive database of Indian medicinal plants for diabetes. *Bioinformation* 2013;**9**:378–80.
53. Ashfaq UA, Mumtaz A, Qamar TU, *et al.* MAPS database: medicinal plant activities, phytochemical and structural database. *Bioinformation* 2013;**9**:993–5.
54. Meetei PA, Singh P, Nongdam P, *et al.* NeMedPlant: a database of therapeutic applications and chemical constituents of medicinal plants from north-east region of India. *Bioinformation* 2012;**8**:209–11.
55. Zeng X, Zhang P, He W, *et al.* NPASS: natural product activity and species source database for natural product research discovery and tool development. *Nucleic Acids Res* 2018;**46**:D1217–22.
56. Zeng X, Zhang P, Wang Y, *et al.* CMAUP: a database of collective molecular activities of useful plants. *Nucleic Acids Res* 2019;**47**:D1118–27.
57. James P, Mathai VA, Shajikumar S, *et al.* DIACAN: integrated database for antidiabetic and anticancer medicinal plants. *Bioinformation* 2013;**9**:941–3.
58. Al-Zahrani AA. Saudi anti-human cancer plants database (SACPD): a collection of plants with anti-human cancer activities. *Oncol Rev* 2018;**12**:349.
59. Hu R, Ren G, Sun G, *et al.* TarNet: an evidence-based database for natural medicine research. *PLoS One* 2016;**11**:e0157222.
60. Nakamura Y, Afendi FM, Parvin AK, *et al.* KNApSAcK metabolite activity database for retrieving the relationships between metabolites and biological activities. *Plant Cell Physiol* 2014;**55**:e7.
61. Fang YC, Huang HC, Chen HH, *et al.* TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining. *BMC Complement Altern Med* 2008;**8**:58.
62. Ye H, Ye L, Kang H, *et al.* HIT: linking herbal active ingredients to targets. *Nucleic Acids Res* 2011;**39**:D1055–9.
63. Boyle B, Hopkins N, Lu Z, *et al.* The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics* 2013; **14**:16.
64. Schneider CA, Rasband WS, Eliceiri KWNIH. Image to ImageJ: 25 years of image analysis. *Nat Methods* 2012;**9**: 671–5.
65. Hartmann A, Czauderna T, Hoffmann R, *et al.* HTPheno: an image analysis pipeline for high-throughput plant phenotyping. *BMC Bioinformatics* 2011;**12**:148.
66. Fahlgren N, Feldman M, Gehan MA, *et al.* A versatile Phenotyping system and analytics platform reveals diverse temporal responses to water availability in Setaria. *Mol Plant* 2015;**8**:1520–35.
67. Gehan MA, Fahlgren N, Abbasi A, *et al.* PlantCV v2: image analysis software for high-throughput plant phenotyping. *PeerJ* 2017;**5**:e4088.
68. Higo K, Ugawa Y, Iwamoto M, *et al.* Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res* 1999;**27**:297–300.
69. Crooks GE, Hon G, Chandonia JM, *et al.* WebLogo: a sequence logo generator. *Genome Res* 2004;**14**:1188–90.
70. Lescot M, Dehais P, Thijs G, *et al.* PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res* 2002;**30**:325–7.
71. Hofacker IL. Vienna RNA secondary structure server. *Nucleic Acids Res* 2003;**31**:3429–31.
72. Janssen S, Giegerich R. The RNA shapes studio. *Bioinformatics* 2015;**31**:423–5.
73. Langmead B, Trapnell C, Pop M, *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**:R25.
74. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods* 2012;**9**:357–9.
75. Grabherr MG, Haas BJ, Yassour M, *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011;**29**:644–52.
76. Xie YL, Wu GX, Tang JB, *et al.* SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 2014;**30**:1660–6.
77. Trapnell C, Roberts A, Goff L, *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat Protoc* 2012;**7**:562–78.
78. Pertea M, Kim D, Pertea GM, *et al.* Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* 2016;**11**:1650–67.
79. Zhai J, Song J, Cheng Q, *et al.* PEA: an integrated R toolkit for plant epitranscriptome analysis. *Bioinformatics* 2018;**34**:3747–9.
80. Wu B. Differential gene expression detection using penalized linear regression models: the improved SAM statistics. *Bioinformatics* 2005;**21**:1565–71.
81. An J, Lai J, Sajjanhar A, *et al.* miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data. *BMC Bioinformatics* 2014;**15**:275.
82. Yang X, Li L. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics* 2011;**27**:2614–5.
83. Yu D, Wan Y, Ito H, *et al.* PmiRDiscVali: an integrated pipeline for plant microRNA discovery and validation. *BMC Genomics* 2019;**20**:133.
84. Fahlgren N, Hill ST, Carrington JC, *et al.* P-SAMS: a web site for plant artificial microRNA and synthetic trans-acting small interfering RNA design. *Bioinformatics* 2016;**32**:157–8.
85. Yu D, Meng Y, Zuo Z, *et al.* NATpipe: an integrative pipeline for systematical discovery of natural antisense transcripts (NATs) and phase-distributed nat-siRNAs from de novo assembled transcriptomes. *Sci Rep* 2016;**6**:21666.
86. Singh U, Khemka N, Rajkumar MS, *et al.* PLncPRO for prediction of long non-coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive lncRNAs in rice and chickpea. *Nucleic Acids Res* 2017;**45**:e183.
87. Chen L, Yu Y, Zhang X, *et al.* PcircRNA_finder: a software for circRNA prediction in plants. *Bioinformatics* 2016;**32**:3528–9.
88. Dai X, Zhuang Z, Zhao PX. psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res* 2018;**46**:W49–54.

89. Bonnet E, He Y, Billiau K, *et al*. TAPIR, a web server for the prediction of plant microRNA targets including target mimics. *Bioinformatics* 2010;**26**:1566–8.
90. Addo-Quaye C, Miller W, Axtell MJ. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* 2009;**25**:130–1.
91. Kakrana A, Hammond R, Patel P, *et al*. sPARTA: a parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. *Nucleic Acids Res* 2014;**42**:e139.
92. Mal C, Aftabuddin M, IIKmTA KS. Inter and intra kingdom miRNA-target analyzer. *Interdiscip Sci* 2018;**10**:538–43.
93. Dai X, Sinharoy S, Udvardi M, *et al*. PlantTFcat: an online plant transcription factor and transcriptional regulator categorization and analysis tool. *BMC Bioinformatics* 2013;**14**:321.
94. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
95. Tian T, Liu Y, Yan H, *et al*. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res* 2017;**45**:W122–9.
96. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559.
97. Marwah VS, Kinaret PAS, Serra A, *et al*. INfORM: inference of network response modules. *Bioinformatics* 2018;**34**: 2136–8.
98. Moriya Y, Shigemizu D, Hattori M, *et al*. PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res* 2010;**38**:W138–43.
99. Rohart F, Gautier B, Singh A, *et al*. mixOmics: an R package for ’omics feature selection and multiple data integration. *PLoS Comput Biol* 2017;**13**:e1005752.
100. Pan F, Kamath K, Zhang K, *et al*. Integrative Array Analyzer: a software package for analysis of cross-platform and cross-species microarray data. *Bioinformatics* 2006;**22**:1665–7.
101. Wright GW, Simon RM. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* 2003;**19**:2448–55.
102. Sharma V, Sarkar IN. Bioinformatics opportunities for identification and study of medicinal plants. *Brief Bioinform* 2013;**14**:238–50.
103. Furbank RT, Tester M. Phenomics–technologies to relieve the phenotyping bottleneck. *Trends Plant Sci* 2011;**16**: 635–44.
104. Fiorani F, Schurr U. Future scenarios for plant phenotyping. *Annu Rev Plant Biol* 2013;**64**:267–91.
105. Fahlgren N, Gehan MA, Baxter I. Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. *Curr Opin Plant Biol* 2015;**24**:93–9.
106. Rahaman MM, Chen D, Gillani Z, *et al*. Advanced phenotyping and phenotype data analysis for the study of plant growth and development. *Front Plant Sci* 2015;**6**:619.
107. Fraas S, Luthen H. Novel imaging-based phenotyping strategies for dissecting crosstalk in plant development. *J Exp Bot* 2015;**66**:4947–55.
108. Kuijken RC, van Eeuwijk FA, Marcelis LF, *et al*. Root phenotyping: from component trait in the lab to breeding. *J Exp Bot* 2015;**66**:5389–401.
109. Singh A, Ganapathysubramanian B, Singh AK, *et al*. Machine learning for high-throughput stress phenotyping in plants. *Trends Plant Sci* 2016;**21**:110–24.
110. Salon C, Avice JC, Colombie S, *et al*. Fluxomics links cellular functional analyses to whole-plant phenotyping. *J Exp Bot* 2017;**68**:2083–98.
111. Ghanem ME, Marrou H, Sinclair TR. Physiological phenotyping of plants for crop improvement. *Trends Plant Sci* 2015;**20**:139–44.
112. Perez-Riverol Y, Gatto L, Wang R, *et al*. Ten simple rules for taking advantage of git and GitHub. *PLoS Comput Biol* 2016;**12**:e1004947.
113. Kim NH, Jayakodi M, Lee SC, *et al*. Genome and evolution of the shade-requiring medicinal herb *Panax ginseng*. *Plant Biotechnol J* 2018;**16**:1904–17.
114. Kang SH, Lee JH, Lee HO, *et al*. Complete chloroplast genome and 45S nrDNA sequences of the medicinal plant species *Glycyrrhiza glabra* and *Glycyrrhiza uralensis*. *Genes Genet Syst* 2018;**93**:83–9.
115. Shen X, Wu M, Liao B, *et al*. Complete chloroplast genome sequence and phylogenetic analysis of the medicinal plant *Artemisia annua*. *Molecules* 2017;**22**.
116. Wang W, Yu H, Wang J, *et al*. The complete chloroplast genome sequences of the medicinal plant *Forsythia suspensa* (Oleaceae). *Int J Mol Sci* 2017;**18**.
117. Wu ML, Li Q, Xu J, *et al*. Complete chloroplast genome of the medicinal plant *Amomum compactum*: gene organization, comparative analysis and phylogenetic relationships within Zingiberales. *Chin Med* 2018;**13**:10.
118. Xie Q, Shen KN, Hao X, *et al*. The complete chloroplast genome of Tianshan snow lotus (*Saussurea involucrata*), a famous traditional Chinese medicinal plant of the family Asteraceae. *Mitochondrial DNA A DNA Mapp Seq Anal* 2017;**28**:294–5.
119. Yang Z, Huang Y, An W, *et al*. Sequencing and structural analysis of the complete chloroplast genome of the medicinal plant *Lycium chinense* mill. *Plants (Basel)* 2019;**8**.
120. Zhang C, Liu T, Yuan X, *et al*. The plastid genome and its implications in barcoding specific-chemotypes of the medicinal herb *Pogostemon cablin* in China. *PLoS One* 2019;**e0215512**:14.
121. Jamali SH, Cockram J, Hickey LT. Insights into deployment of DNA markers in plant variety protection and registration. *Theor Appl Genet* 2019.
122. Mishra P, Kumar A, Nagireddy A, *et al*. DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market. *Plant Biotechnol J* 2016;**14**:8–21.
123. Heubl G. New aspects of DNA-based authentication of Chinese medicinal plants by molecular biological techniques. *Planta Med* 2010;**76**:1963–74.
124. Li B, Cui G, Shen G, *et al*. Targeted mutagenesis in the medicinal plant *Salvia miltiorrhiza*. *Sci Rep* 2017;**7**:43320.
125. Zhao MM, Zhang G, Zhang DW, *et al*. ESTs analysis reveals putative genes involved in symbiotic seed germination in *Dendrobium officinale*. *PLoS One* 2013;**8**:e72705.
126. Kiyama RDNA. Microarray-based screening and characterization of traditional Chinese medicine. *Microarrays* 2017;**6**.
127. Amini H, Naghavi MR, Shen T, *et al*. Tissue-specific transcriptome analysis reveals candidate genes for terpenoid and phenylpropanoid metabolism in the medicinal plant *Ferula assafoetida*. *G3 (Bethesda)* 2019;**9**:807–16.
128. Hao D, Ma P, Mu J, *et al*. De novo characterization of the root transcriptome of a traditional Chinese medicinal plant *Polygonum cuspidatum*. *Sci China Life Sci* 2012;**55**:452–66.
129. Vashisht I, Pal T, Sood H, *et al*. Comparative transcriptome analysis in different tissues of a medicinal herb, *Picrorhiza*

kurroa pinpoints transcription factors regulating picrosides biosynthesis. *Mol Biol Rep* 2016;**43**:1395–409.

130. Gao F, Wang J, Wei S, *et al.* Transcriptomic analysis of drought stress responses in *Ammopiptanthus mongolicus* leaves using the RNA-Seq technique. *PLoS One* 2015;**10**:e0124382.
131. Higashi Y, Saito K. Network analysis for gene discovery in plant-specialized metabolism. *Plant Cell Environ* 2013;**36**:1597–606.
132. Sun L, Rai A, Rai M, *et al.* Comparative transcriptome analyses of three medicinal *Forsythia* species and prediction of candidate genes involved in secondary metabolisms. *J Nat Med* 2018;**72**:867–81.
133. She J, Yan H, Yang J, *et al.* croFGD: *Catharanthus roseus* functional genomics database. *Front Genet* 2019;**10**:238.
134. Fan R, Li Y, Li C, *et al.* Differential microRNA analysis of glandular trichomes and young leaves in *Xanthium strumarium* L. reveals their putative roles in regulating terpenoid biosynthesis. *PLoS One* 2015;**10**:e0139002.
135. Khaldun AB, Huang W, Liao S, *et al.* Identification of microRNAs and target genes in the fruit and shoot tip of *Lycium chinense*: a traditional Chinese medicinal plant. *PLoS One* 2015;**10**:e0116334.
136. Zhang M, Dong Y, Nie L, *et al.* High-throughput sequencing reveals miRNA effects on the primary and secondary production properties in long-term subcultured Taxus cells. *Front Plant Sci* 2015;**6**:604.
137. Gupta OP, Karkute SG, Banerjee S, *et al.* Contemporary understanding of miRNA-based regulation of secondary metabolites biosynthesis in plants. *Front Plant Sci* 2017;**8**:374.
138. Xie W, Weng A, Melzig MF. MicroRNAs as new bioactive components in medicinal plants. *Planta Med* 2016;**82**:1153–62.
139. Chin AR, Fong MY, Somlo G, *et al.* Cross-kingdom inhibition of breast cancer growth by plant miR159. *Cell Res* 2016;**26**:217–28.
140. Kumar D, Kumar S, Ayachit G, *et al.* Cross-kingdom regulation of putative miRNAs derived from happy tree in cancer pathway: a systems biology approach. *Int J Mol Sci* 2017;**18**.
141. Hong M, Wang N, Tan HY, *et al.* MicroRNAs and Chinese medicinal herbs: new possibilities in cancer therapy. *Cancers (Basel)* 2015;**7**:1643–57.
142. Zhou Z, Li X, Liu J, *et al.* Honeysuckle-encoded atypical microRNA2911 directly targets influenza a viruses. *Cell Res* 2015;**25**:39–49.
143. Zhang X, Allan AC, Li C, *et al.* De novo assembly and characterization of the transcriptome of the Chinese medicinal herb *Gentiana rigescens*. *Int J Mol Sci* 2015;**16**:11550–73.
144. Salimi V, Maroufi A, Majdi M. Differential expression of 3 beta-HSD and mlncRNAs in response to abiotic stresses in *Digitalis nervosa*. *Cell Mol Biol* 2018;**64**:89–95.
145. Wang M, Wu B, Chen C, *et al.* Identification of mRNA-like non-coding RNAs and validation of a mighty one named MAR in *Panax ginseng*. *J Integr Plant Biol* 2015;**57**:256–70.
146. Jacobs DI, Gaspari M, van der Greef J, *et al.* Proteome analysis of the medicinal plant *Catharanthus roseus*. *Planta* 2005;**221**:690–704.
147. Chen J, Liu SS, Kohler A, *et al.* iTRAQ and RNA-Seq analyses provide new insights into regulation mechanism of symbiotic germination of *Dendrobium officinale* seeds (Orchidaceae). *J Proteome Res* 2017;**16**:2174–87.
148. Chandra DN, Prasanth GK, Singh N, *et al.* Identification of a novel and potent inhibitor of phospholipase A(2) in a medicinal plant: crystal structure at 1.93A and surface Plasmon resonance analysis of phospholipase A(2) complexed with berberine. *Biochim Biophys Acta* 2011;**1814**:657–63.
149. Singh G, Tripathi S, Shanker K, *et al.* Cadmium-induced conformational changes in type 2 metallothionein of medicinal plant *Coptis japonica*: insights from molecular dynamics studies of apo, partially and fully metalated forms. *J Biomol Struct Dyn* 2019;**37**:1520–33.
150. Upadhyay AK, Sowdhamini R. Genome-wide analysis of domain-swap predicted products in the genome of anti-stress medicinal plant: *Ocimum tenuiflorum*. *Bioinform Biol Insights* 2019;**13**:1177932218821362.
151. Liu K, Yuan C, Li H, *et al.* A qualitative proteome-wide lysine crotonylation profiling of papaya (*Carica papaya* L.). *Sci Rep* 2018;**8**:8230.
152. Shen CJ, Xue J, Sun T, *et al.* Succinyl-proteome profiling of a high taxol containing hybrid Taxus species (Taxus x media) revealed involvement of succinylation in multiple metabolic pathways. *Sci Rep* 2016;**6**.
153. Braun P, Aubourg S, Van Leene J, *et al.* Plant protein interactomes. *Annu Rev Plant Biol* 2013;**64**:161–87.
154. Saito K, Matsuda F. Metabolomics for functional genomics, systems biology and biotechnology. *Annu Rev Plant Biol* 2010;**61**:463–89.
155. Gad HA, El-Ahmady SH, Abou-Shoer MI, *et al.* Application of chemometrics in authentication of herbal medicines: a review. *Phytochem Anal* 2013;**24**:1–24.
156. Ning Z, Lu C, Zhang Y, *et al.* Application of plant metabonomics in quality assessment for large-scale production of traditional Chinese medicine. *Planta Med* 2013;**79**:897–908.
157. Afzan A, Kasim N, Ismail NH, *et al.* Differentiation of Ficus deltoidea varieties and chemical marker determination by UHPLC-TOFMS metabolomics for establishing quality control criteria of this popular Malaysian medicinal herb. *Metabolomics* 2019;**15**:35.
158. Okuda S, Yamada T, Hamajima M, *et al.* KEGG atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res* 2008;**36**:W423–6.
159. Govindaraghavan S, Hennell JR, Sucher NJ. From classical taxonomy to genome and metabolome: towards comprehensive quality standards for medicinal herb raw materials and extracts. *Fitoterapia* 2012;**83**:979–88.
160. Kim SW, Gupta R, Lee SH, *et al.* An integrated biochemical, proteomics, and metabolomics approach for supporting medicinal value of *Panax ginseng* fruits. *Front Plant Sci* 2016;**7**:994.
161. Chen Q, Li M, Wang C, *et al.* Combining targeted metabolites analysis and transcriptomics to reveal chemical composition difference and underlying transcriptional regulation in Maca (*Lepidium meyenii* Walp.) ecotypes. *Genes* 2018;**9**.
162. Bhattacharyya P, Kumaria S, Kumar S, *et al.* Start codon targeted (SCoT) marker reveals genetic diversity of *Dendrobium nobile* Lindl an endangered medicinal orchid species. *Gene* 2013;**529**:21–6.
163. Feng S, He R, Yang S, *et al.* Start codon targeted (SCoT) and target region amplification polymorphism (TRAP) for evaluating the genetic relationship of Dendrobium species. *Gene* 2015;**567**:182–8.
164. Lu JJ, Suo NN, Hu X, *et al.* Development and characterization of 110 novel EST-SSR markers for *Dendrobium officinale* (Orchidaceae). *Am J Bot* 2012;**99**:e415–20.
165. Yan L, Wang X, Liu H, *et al.* The genome of *Dendrobium officinale* illuminates the biology of the important traditional Chinese orchid herb. *Mol Plant* 2015;**8**:922–34.

166. Zhang GQ, Xu Q, Bian C, *et al*. The *Dendrobium catenatum* Lindl. Genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. *Sci Rep* 2016;**6**:19029.

167. Lau SE, Schwarzacher T, Othman RY, *et al*. dsRNA silencing of an R2R3-MYB transcription factor affects flower cell shape in a Dendrobium hybrid. *BMC Plant Biol* 2015;**15**:194.

168. Kui L, Chen H, Zhang W, *et al*. Building a genetic manipulation tool box for orchid biology: Identification of constitutive promoters and application of CRISPR/Cas9 in the orchid, *Dendrobium officinale*. *Front Plant Sci* 2016;**7**:2036.

169. He C, Wu K, Zhang J, *et al*. Cytochemical localization of polysaccharides in *Dendrobium officinale* and the involvement of DoCSLA6 in the synthesis of Mannan polysaccharides. *Front Plant Sci* 2017;**8**:173.

170. Meng C, Zeleznik OA, Thallinger GG, *et al*. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform* 2016;**17**:628–41.