D976–D982 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 26 October 2018*
doi: 10.1093/nar/gky987

ETCM: an encyclopaedia of traditional Chinese medicine

Hai-Yu Xu[1,2,*,†], Yan-Qiong Zhang[1,†], Zhen-Ming Liu[3,†], Tong Chen[2], Chuan-Yu Lv[3], Shi-Huan Tang[1], Xiao-Bo Zhang[2], Wei Zhang[1], Zhi-Yong Li[4], Rong-Rong Zhou[4], Hong-Jun Yang[1,*], Xiu-Jie Wang[5,*] and Lu-Qi Huang[2,*]

[1]Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China, [2]National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China, [3]State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences, Peking University, Beijing 100191, China, [4]China Minority Traditional Medical Center, Minzu University of China, Beijing 100081, China and [5]Key Laboratory of Genetic Networks, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

Received August 15, 2018; Revised September 27, 2018; Editorial Decision October 07, 2018; Accepted October 20, 2018

ABSTRACT

Traditional Chinese medicine (TCM) is not only an effective solution for primary health care, but also a great resource for drug innovation and discovery. To meet the increasing needs for TCM-related data resources, we developed ETCM, an Encyclopedia of Traditional Chinese Medicine. ETCM includes comprehensive and standardized information for the commonly used herbs and formulas of TCM, as well as their ingredients. The herb basic property and quality control standard, formula composition, ingredient drug-likeness, as well as many other information provided by ETCM can serve as a convenient resource for users to obtain thorough information about a herb or a formula. To facilitate functional and mechanistic studies of TCM, ETCM provides predicted target genes of TCM ingredients, herbs, and formulas, according to the chemical fingerprint similarity between TCM ingredients and known drugs. A systematic analysis function is also developed in ETCM, which allows users to explore the relationships or build networks among TCM herbs, formulas,ingredients, gene targets, and related pathways or diseases. ETCM is freely accessible at http://www.nrc.ac.cn:9090/ETCM/. We expect ETCM to develop into a major data warehouse for TCM and to promote TCM related researches and drug development in the future.

INTRODUCTION

Traditional Chinese medicine (TCM) holds great potentials for health improvement as well as prevention and treatment of various diseases, especially complex diseases such as autoimmune disorders, cardiovascular diseases and cancers (1–3). TCM is also a great resource for modern drug research and development. Many TCM-derived drugs have shown remarkable effects in curing diseases, such as artemisinin, digitoxin, quinine and celastrol. The most recognized effect of TCM is the use of artemisinin-based remedies to treat malaria, which was awarded the Nobel Prize in Physiology and Medicine in 2015 (4). After that, growing attention has been attracted to TCM, which also brought the increasing needs for TCM related data resources.

Herbs are the most commonly used substances in TCM. Over 11 000 herb plants have been recorded in various TCM related pharmacopeia, and the commonly used ones are ∼700 species. As TCM usually combines multiple herbs as formulas in disease treatments, thousands of herbal formulas have been generated and widely applied in clinics. In theory, TCM herbal formulas contain multiple effective ingredients, thus can simultaneously regulate many targets within cells, therefore to reestablish balanced physiological regulatory networks of patients and to cure diseases (5,6). Yet the molecular targets of most TCM formulas and ingredients still remain elusive, which becomes one of the biggest hurdles in the application of TCM and TCM-based drug discovery.

The recent development of several TCM-related databases, such as HIT (7), TCMGeneDIT (8), TCMMESH (9), TCM-ID (10), TCMSP (11), TCMID (12,13),

*To whom correspondence should be addressed. Tel: +86 10 64014411; Fax: +86 10 64013996; Email: hyxu@icmm.ac.cn
Correspondence may also be addressed to Lu-Qi Huang. Tel: +86 10 64014411; Fax: +86 10 64013996; Email: huangluqi01@126.com
Correspondence may also be addressed to Xiu-Jie Wang. Tel: +86 10 64806590; Fax: +86 10 64806595; Email: xjwang@genetics.ac.cn
Correspondence may also be addressed to Hong-Jun Yang. Tel: +86 10 64014411; Fax: +86 10 64013996; Email: hjyang@icmm.ac.cn
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

have provided useful data and tools for TCM-based research and drug discovery. Yet most databases only focus on herbs and their components, thus the relationships between formulas and herbs/components are missing. None of these databases provide the habitat and quality control information of herbs, which is considered as a major factor for the effectiveness of TCM. In addition, the design of some databases are obsolete, with limited information and are difficult to use. Here, we present ETCM, an Encyclopedia of Traditional Chinese Medicine (http://www.nrc.ac.cn:9090/ETCM/), which includes multiple aspects of clinical and functional essential information on 403 TCM herb species, 3962 TCM formulas, 7274 herbal ingredients, 2266 validated or predicted drug targets, as well as 3027 related diseases. All these information is comprehensively linked to each other in the database and displayed with user friendly interfaces, which can serve as a valuable resource for TCM-related research and drug discovery.

DATA RESOURCES AND DATABASE CONTENTS

ETCM combines data from multiple sources and provides comprehensive information on the commonly used TCM herbs, herbal formulas, ingredients, as well as the predicted drug targets of ingredients, and their related diseases. The major contents of ETCM are summarized in Table 1.

The herb information

The herb section contains the general and drug effectiveness-related information of 403 commonly used Chinese herbs, including the name, type, species, collection time, property, flavor, meridian tropism, indication and specification of herbs. These information was collected from the Pharmacopoeia of the People's Republic of China (2015 version). The pictures and habitat map of herbs were obtained from the Fourth National Survey on Chinese Materia Medica Resources (http://www.zyzypc.com.cn/), which was initiated by the State Administration of Traditional Chinese Medicine in China and carried out a large-scale investigation on Chinese medicinal resources in 1345 counties of China. The colored provinces on the habitat maps are locations where the herbs can grow in China. The ingredients of herbs were mainly extracted from published papers, as well as from ChEMBL (https://www.ebi.ac.uk/chembl/) and PubChem (https://pubchem.ncbi.nlm.nih.gov/).

The formula information

A total of 3962 TCM prescriptions approved by China Food and Drug Administration (CFDA) were collected and stored in the formula section of ETCM. These prescriptions are the commonly used TCM formulas for treating various diseases or maintaining healthy body conditions in clinical practice. The information of formula name, type, dosage form, composition, applicable indications and syndromes, as well as route of administration were obtained from the Pharmacopoeia of the People's Republic of China (2015 version) and Drug Standard of Ministry of Public Health of the People's Republic of China (1993 version). The external links to CFDA are provided to search more detailed drug information by clicking the formula name.

Quality standard of herbs and TCM prescriptions

To facilitate quality evaluation of TCM, the quantitative standards of marker ingredients of herbs and formulas are provided in the herb and formula sections of ETCM. The quantitative standards are set according to the Pharmacopoeia of the People's Republic of China (2015 version), which are the official TCM quality evaluation standards in China.

The ingredient information

The ingredients of herbs and formulas reported in ETCM were collected manually according to the Pharmacopoeia of the People's Republic of China (2015 version) and other literatures. For each ingredient, its name, molecular formula, molecular weight, 2D structure, partition coefficient (ALogP), distribution coefficient (LogD), molecular solubility, molecular volume, surface area and polar surface area, as well as the number of rotatable bonds, H acceptors and H donors are included. These physico-chemical properties were calculated using the Pipeline Pilot software (version 7.5). External links to Public database sources, such as ChEMBL (https://www.ebi.ac.uk/chembl/) and PubChem (https://pubchem.ncbi.nlm.nih.gov/), are also provided for each ingredient.

To estimate drug-likeness of each ingredient, pharmacokinetic parameters were calculated based on models in Pipeline Pilot ADMET collection, such as aqueous solubility, blood brain barrier penetration, CYP450 2D6 inhibition, hepatotoxicity, human intestinal absorption and plasma protein binding. A quantitative metrics, known as the quantitative estimate of drug-likeness (QED), was used for assessing drug-likeness, and the estimated values range from 0 (all properties are unfavorable) to 1 (all properties are favorable). The reported mean QED values for attractive and unattractive components in drug development are 0.67 and 0.49, respectively (14–16). Thus, we classified all 7274 ingredients collected in ETCM into three groups according to their QED scores, good (QED $>$ 0.67), moderate (0.49 $\leq$ QED $\leq$ 0.67) and weak (QED $<$ 0.49).

The target information

We predicted potential targets for herbal ingredients in ETCM using MedChem Studio (version 3.0), an efficient drug similarity search tool to identify known drugs with high structural similarity (Tanimoto score $>$ 0.8) to herbal ingredients. The Tanimoto Score is in the range of [0,1], where '0' denotes completely different structures between ingredients and known drugs, and '1' denotes identical structures of two components. The therapeutic targets in the DrugBank database (17) of known drugs are considered as the candidate targets of the herbal ingredients with Tanimoto Scores $>$0.8 to the known drugs. The biological functions and participated pathways of the candidate drug targets are investigated according to the Gene Ontology (18) and KEGG (Kyoto Encyclopedia of Genes and

Table 1. Content statistics of ETCM

Data type	Number	Resources
Herbs	403	The Fourth National Survey on Chinese Materia Medica Resources (http://www.zyzypc.com.cn/)
Herbs with quantitative information of marker ingredients	*263*	Pharmacopoeia of the People's Republic of China (2015 version)
TCM formulas	3962	China Food and Drug Administration (http://eng.sfda.gov.cn/WS03/CL0755/)
Formulas with quantitative information of marker ingredients	*478*	Pharmacopoeia of the People's Republic of China (2015 version)
Ingredients	7274	Manual literature retrievals and Pubchem (https://pubchem.ncbi.nlm.nih.gov/)
Ingredients with drug-likeness evaluation	*7269*	A quantitative estimate model of drug-likeness reported byBickerton group [*Nat Chem. 2012 Jan 24;4(2):90-8.*]
Drug target genes	2266	MedChem Studio (version 3.0; Simulations Plus, Inc., Lancaster, CA, USA, 2012)
TCM-related diseases	3027	Human Phenotype Ontology (HPO, http://human-phenotype-ontology.github.io/) Online Mendelian Inheritance in Man (OMIM, http://www.omim.org/) Database of gene-disease associations (DisGeNET, http://www.disgenet.org/web/DisGeNET/menu;jsessionid=c807m1cvnhyn1uc0dr111ox8c, v5.0) ORPHANET (https://www.orpha.net/consor/cgi-bin/Disease.php?lng=EN)

Genomes) pathway database (19). Enrichment of Gene Ontology terms and KEGG pathways by target genes of each formula or herb is analyzed using in house python scripts with hypergeometric test.

The disease information

The disease section records the detailed information of diseases and their related genes, which were collected from Human Phenotype Ontology (HPO, March 2018 release), Online Mendelian Inheritance in Man (OMIM, April 2018 release), Database of gene–disease associations (DisGeNET, v5.0) and ORPHANET database. The inconsistent gene or protein IDs of different resources were manually inspected and converted into Official Gene Symbols and UniProt Accession Numbers. The diseases are linked to herbs and formulas according to the overlap of disease-causing genes and the putative target genes of herbs/formulas.

Network analysis

ETCM uses the network module of a dynamic browser based visualization library vis.js (v4.21.0, http://visjs.org/index.html) to construct user-friendly networks for multi-level interactions among herbs/formulas, targets and diseases. The gene-gene interaction data were collected from five existing molecular interaction databases, including Reactome (https://reactome.org/, version 65), Human Protein Reference Database (HPRD, http://www.hprd.org/, Release 9), the Molecular INTeraction Database (MINT, https://mint.bio.uniroma2.it/, updated in August 2011), IntAct Molecular Interaction Database (https://www.ebi.ac.uk/intact/, version: 4.2.12) and Database of Interacting Proteins (DIP, https://dip.doe-mbi.ucla.edu/dip/, updated in Feb 13, 2017). Connections among herbs, formulas, pathways and diseases are established by the ingredients of herbs/formulas and the putative targets of each ingredient.

DATABASE FEATURES AND WEB INTERFACE

The ETCM database provides an easy-to-use platform for users to browse, search, and analyze TCM related information from various aspects.The major functional schema of ETCM is summarized in Figure 1. To facilitate TCM related basic researches, clinical applications and drug development, the ETCM database not only includes detailed medicinal properties of TCM herbs and formulas, but also provides potential links between TCM with target genes and modern diseases.

TCM-specific functional features

TCM understands human bodies and diseases from a special systematic point of view, which is quite different from modern anatomy and medicines. For example, in TCM, herbs are classified according to their flavours (sour, salty, sweet, bitter and pungent), the properties of herbs are characterized according to their effects (cold, hot, warm, cool, and even), and the target organs of herbs are defined by meridian tropisms (lung meridian, liver meridian, etc.). All these information are presented in the 'Herbs' section of ETCM (Figure 2A). By clicking on the pie chart of each above mentioned category, users can obtain a full list of all herbs belonging to each category (Figure 2B).

Detailed information of each herb can be retrieved by clicking on its Chinese or Pinyin name, including the habitat, best collection time, property, flavor, meridian tropism, indications, and chemical components of the herb. The picture of each herb as well as its habitat map in China and

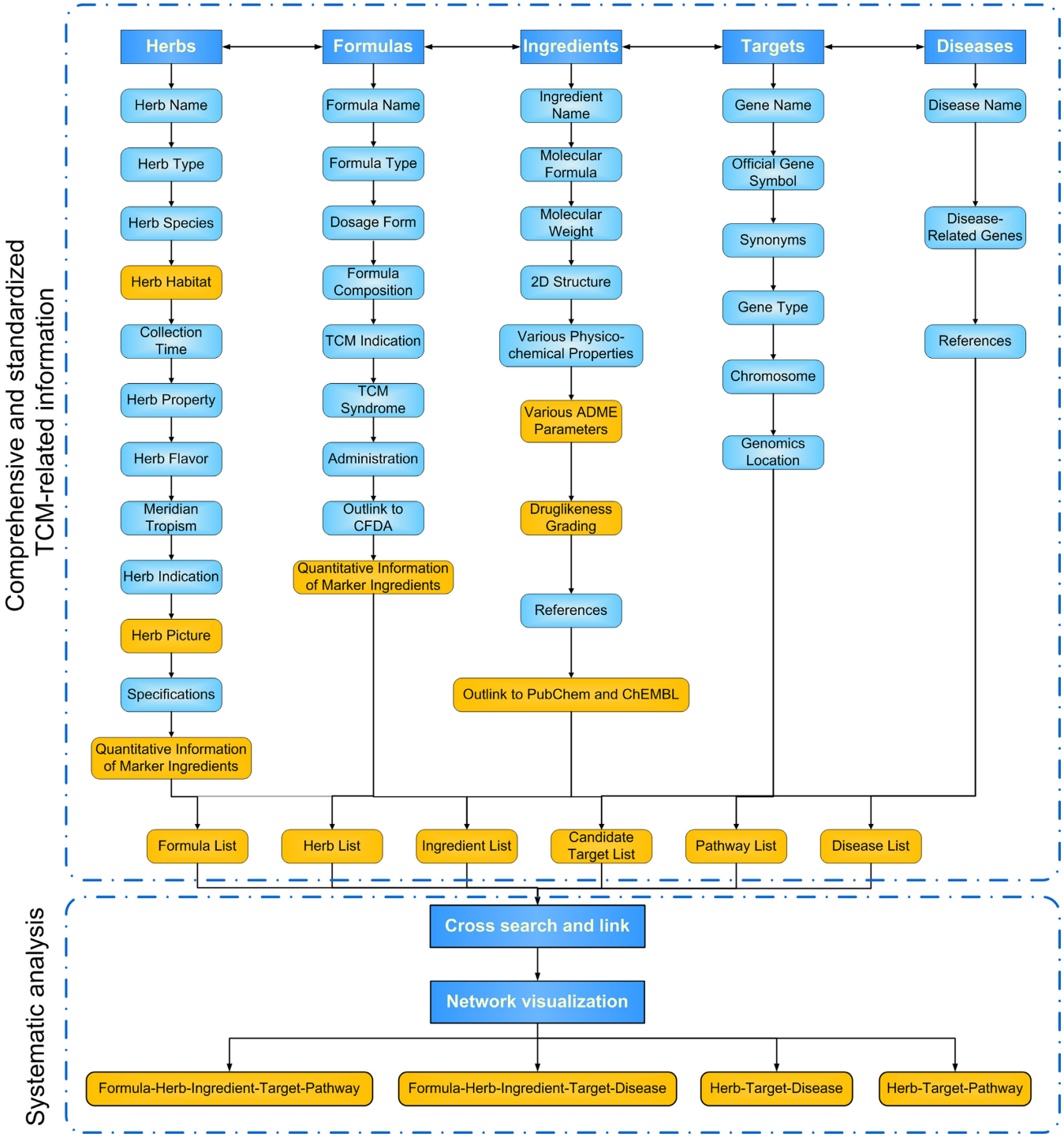

Figure 1. The schema and objectives of ETCM. Boxes with yellow highlights refer to the novel information and functions in ETCM compared with other available TCM-related data resources.

quality control standard are provided (Figure 2C and D). Formulas containing each herb are also provided in the herb information page, clicking each formula name can link to the information page of the formula. It is worth to note that the herb indication information provided in the 'Herbs' section is described according to the records in the Pharmacopoeia of the People's Republic of China (2015 version), these indications are different from modern diseases, therefore we tried to use genes related to both TCM ingredients and modern diseases to build the links between TCM herb indications and modern diseases. Similarly, formulas collected in the ETCM database are also classified according to TCM syndromes, such as heart-clearing, blood-regulation, etc.

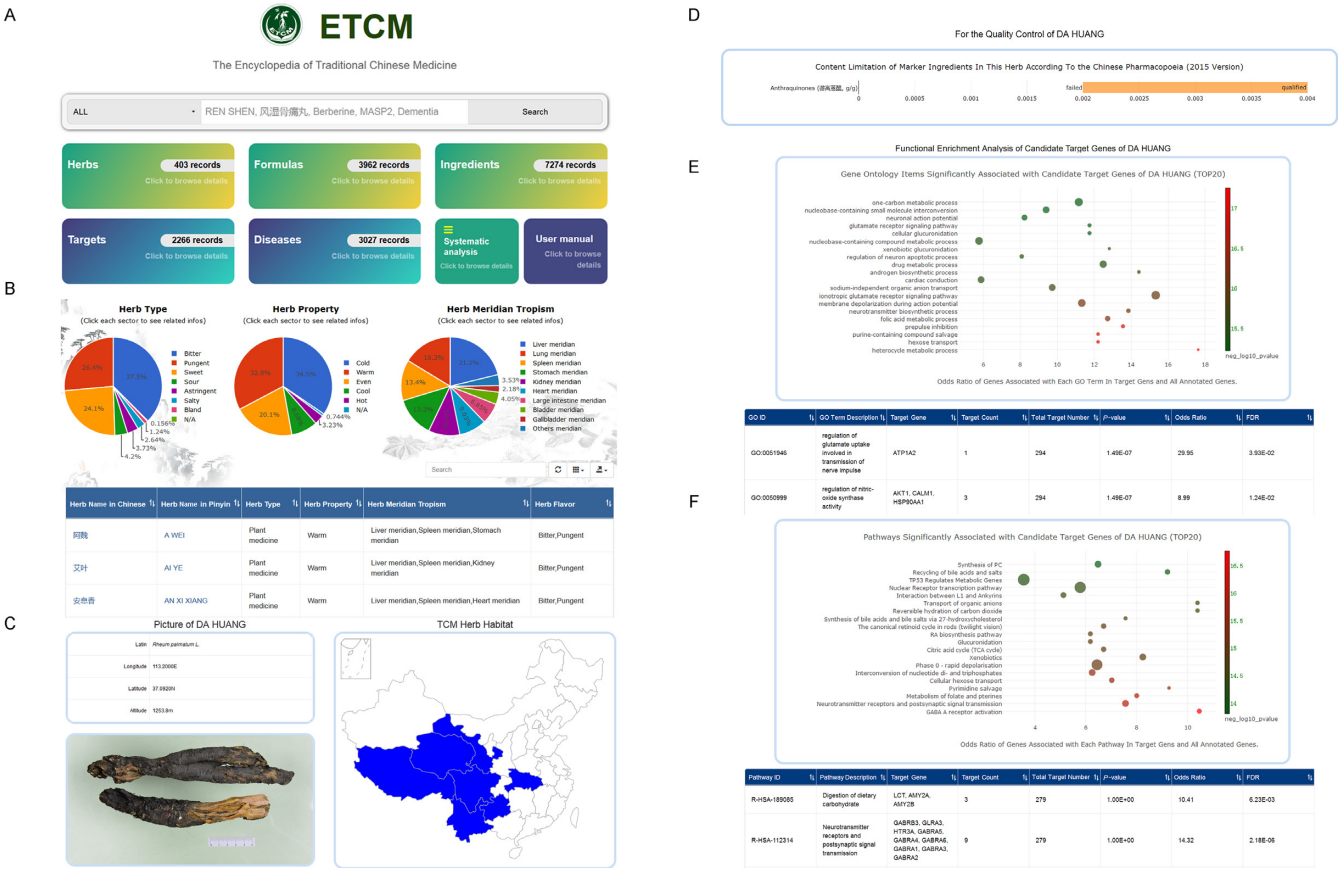

Figure 2. Major function illustration of ETCM. (**A**) Function summary of ETCM. (**B**) Summary of herbs included in ETCM. Clicking on any proportion of the pie charts will bring out corresponding herbs in the below list. (**C**) The picture and habitat map of herb DA HUANG. (**D**) Quality control standard of herb DA HUANG. (**E**) Enrichment and list of GO terms of the putative target genes of herb DA HUANG. (**F**) Pathways enriched by the putative target genes of herb DA HUANG.

Links to molecular biology and modern diseases

To facilitate the understanding of TCM functions from the modern science point of view, we performed target prediction for ingredients of the collected herbs and formulas according to the 2D MACCS fingerprint similarity between TCM ingredients and known drugs in the DrugBank database, and considered the target genes of known drugs as the putative targets of TCM ingredients passed the fingerprint similarity test of the known drugs. The collective targets of all ingredients of a herb or formula are considered as the putative targets of the herb or formula. Diseases associated with those target genes are also considered as diseases that may be cured by the herbs or formulas. Gene Ontology terms or pathways enriched by genes targeted by certain ingredients, herbs, formulas or associated with certain diseases, are also included in ETCM (Figure 2E and F).

To better illustrate the relationships among ingredients, herbs, formulas, target genes, gene-involved pathways and diseases, ETCM provides a systematic analysis function, which allows users to build networks among two or more above mentioned items. By entering a query item and selecting one or more categories, users are capable to build herb-ingredient-target, ingredient-gene-pathway-disease, as well as many other networks (Figure 3). They can also mark or modify nodes and edges of a network to facilitate further researchers. All database contents and analysis results are available for download, and a clear user-manual is also provided.

DISCUSSION

TCM regards human body as a whole system and gives specific prescriptions to each individual patient according to their disease conditions, these concepts are in good accordance with the goals of precision medicine. In recent years, the values of TCM are more and more appreciated. With the hundreds and thousands of historically practiced herbs and formulas, TCM can also serve as a treasure house for modern drug development. Yet TCM-related data resources are still very limited. To meet the increasing needs, we presented ETCM as a novel TCM-related database, which includes various aspects of essential information related to TCM herbs and formulas. Compared with the limited available TCM-related data resources (7–13), ETCM is more comprehensive (e.g. with the habitat and quality control information of herbs, and drug-likeness information of ingredients) and more functional versatile (e.g. allows users to carry out cross-sectional analysis and build networks), as well as includes more items in each category.

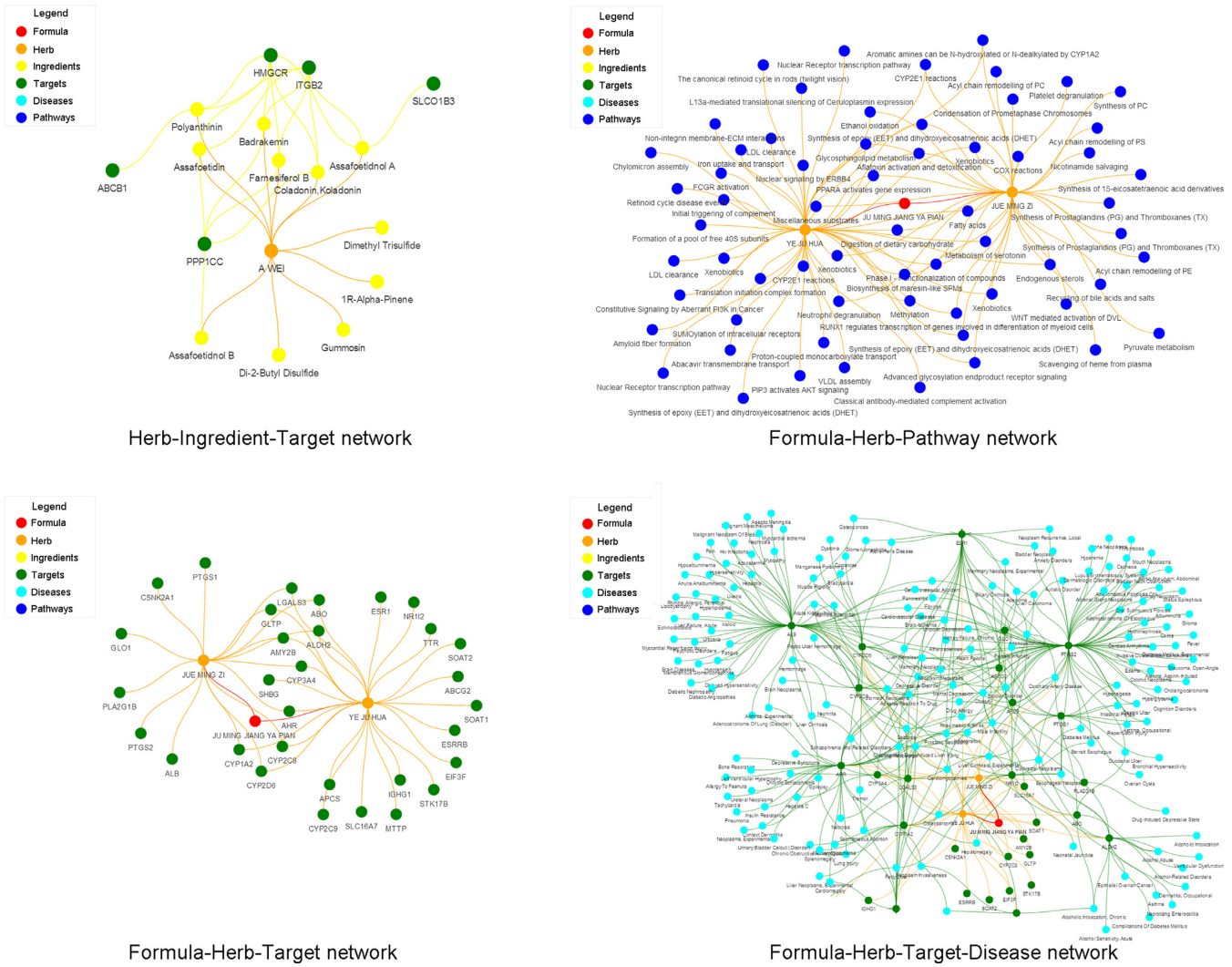

Figure 3. Examples of various networks constructed by the systematic analysis function of ETCM.

To facilitate mechanistic studies on TCM herbs and formulas, we predicted target genes of TCM ingredients according to the structural and chemical similarity of ingredients with known drugs. Although the prediction method we used has been evaluated as one of the best performed methods in similarity based drug discovery (20,21), there may still be many false positives in the prediction results. Thus, the target prediction results could only serve as a mechanism indication of TCM ingredients, herbs and formulas, and await to be investigated in the future.

Multi-drug combination therapeutics has been considered as a rational and efficient therapeutic approach to control complex diseases by regulating various targets and exerting multiple pharmacological effects simultaneously (22). As an empirical system of multi-drug combination therapeutics, TCM prescriptions often include two or more Chinese herbs, the combinational effect of which may be greater than the sum of the individual effects (23). Growing lines of evidence have shown that TCM patent prescriptions with multi-herb combination can increase the possibility of conquering complex diseases with reduced side effects and less adaptive resistance (24). Therefore, understanding the synergistic mechanisms of herbs contained in TCM prescriptions may be of great significance to optimize and discover novel drug combinations. The ETCM database not only includes detailed information of 3962 CFDA approved TCM prescriptions which have been extensively used in clinics with verified therapeutic effects, but also provides a systematic analysis tool for users to investigate putative mechanisms underlying the synergistic effects of ingredients or herbs in a TCM prescription.

In conclusion, ETCM is a comprehensive data resource for aiding the mechanistic investigation, new drug discovery, and clinical application of TCM. It is free for academic use and the data can be conveniently exported. The database will be continually updated and expanded to include new data and functions in the future.

DATA AVAILABILITY

All data of ETCM is available at http://www.nrc.ac.cn:9090/ETCM/.

ACKNOWLEDGEMENTS

We thank Mr. Pu XUE and Mr. Moyu LIU in EHBIO Gene Technology (Beijing) Co., Ltd for their help on the construction of the ETCM database.

FUNDING

Key project at the National Natural Science Foundation of China [81830111 to H.-Y.X., 81330086 to H.-J.Y., 81473414 to H.-Y.X., 81673834 to Y.-Q. Z., 21772005 to Z.M.L.]; Key project at central government level [2060302 to L.-Q.H.]; 973 Program of China [2015CB554406 to H.-J.Y.]; National Key Technology R&D Program of China [2011BAI07B08 to H.-J.Y. and 2016YFC 0903001 to X.-J.W.]; Fundamental Research Funds for the Central public welfare research institutes [L2017018 to Y.-Q.Z.]; National Key Research and Development Program of China [2017YFC1702104 to H.-Y.X., 2017YFC1702303 to W.Z.]. Funding for open access charge: National Natural Science Foundation of China [81830111].

Conflict of interest statement. None declared.

REFERENCES

1. Cheung,F. (2011) TCM: Made in China. *Nature*, **480**, S82–S83.
2. Yan,Z., Lai,Z. and Lin,J. (2017) Anticancer properties of traditional chinese medicine. *Comb. Chem. High Throughput Screen.*, **20**, 423–429.
3. Ma,H.D., Deng,Y.R., Tian,Z. and Lian,Z.X. (2013) Traditional Chinese medicine and immune regulation. *Clin. Rev. Allergy Immunol.*, **44**, 229–241.
4. Su,X.Z. and Miller,L.H. (2015) The discovery of artemisinin and the nobel prize in physiology or medicine. *Sci. China Life Sci.*, **58**, 1175–1179.
5. Sucher,N.J. (2013) The application of Chinese medicine to novel drug discovery. *Expert Opin Drug Discov.*, **8**, 21–34.
6. LI,S. (2015) Mapping ancient remedies: applying a network approach to traditional Chinese medicine. *Sponsored Suppl. Sci.*, **350**, S72.
7. Ye,H., Ye,L., Kang,H., Zhang,D., Tao,L., Tang,K., Liu,X., Zhu,R., Liu,Q., Chen,Y.Z. *et al.* (2011) HIT: linking herbal active ingredientsto targets. *Nucleic Acids Res.*, **39**, D1055–D1059.
8. Fang,Y.C., Huang,H.C. and Chen,HH. (2008) TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining. *BMC Complement. Altern. Med.*, **8**, 58.
9. Zhang,R.Z., Yu,S.J. and Bai,H. (2017) TCM-Mesh: The database and analytical system for network pharmacology analysis for TCM preparations. *Sci. Rep.*, **7**, 2821.
10. Chen,X., Zhou,H., Liu,Y.B., Wang,J.F., Li,H., Ung,C.Y., Han,L.Y., Cao,Z.W. and Chen,Y.Z. (2006) Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation. *Br. J. Pharmacol.*, **149**, 1092–1103.
11. Ru,J., Li,P., Wang,J., Zhou,W., Li,B., Huang,C., Li,P., Guo,Z., Tao,W., Yang,Y. *et al.* (2014) TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. *J. Cheminform.*, **6**, 13.
12. Huang,L., Xie,D., Yu,Y., Liu,H., Shi,Y., Shi,T. and Wen,C. (2018) TCMID 2.0: a comprehensive resource for TCM. *Nucleic Acids Res.*, **46**, D1117–D1120.
13. Xue,R., Fang,Z., Zhang,M., Yi,Z., Wen,C. and Shi,T. (2013) TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis. *Nucleic Acids Res.*, **41**, D1089–D1095.
14. Bickerton,G.R., Paolini,G.V., Besnard,J., Muresan,S. and Hopkins,A.L. (2012) Quantifying the chemical beauty of drugs. *Nat. Chem.*, **4**, 90–98.
15. Doak,B.C., Zheng,J., Dobritzsch,D. and Kihlberg,J. (2016) How Beyond rule of 5 drugs and clinical candidates bind to their targets. *J. Med. Chem.*, **59**, 2312–2327.
16. Doak,B.C., Over,B., Giordanetto,F. and Kihlberg,J. (2014) Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. *Chem. Biol.*, **21**, 1115–1142.
17. Wishart,D.S., Feunang,Y.D., Guo,A.C., Lo,E.J., Marcu,A., Grant,J.R., Sajed,T., Johnson,D., Li,C., Sayeeda,Z. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
18. Gene Ontology Consortium. (2015) Gene ontology consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
19. Kanehisa,M. and Goto,S. (2000) KEGG: kyotoencyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
20. Berenger,F., Vu,O. and Meiler,J. (2017) Consensus queries in ligand-based virtual screening experiments. *J. Cheminform.*, **9**, 60.
21. Huang,Q., Kang,H., Zhang,D.F., Sheng,Z., Liu,Q., Zhu,R.X. and Cao,Z.W. (2011) Comparison of ligand-, target structure-, and protein-ligand interaction Fingerprint-based virtual screening methods. *Acta Chim. Sinica*, **5**, 515–522.
22. Lehár,J., Krueger,A.S., Avery,W., Heilbut,A.M., Johansen,L.M., Price,E.R., Rickles,R.J., Short,G.F., Staunton,J.E., Jin,X. *et al.* (2009) Synergistic drug combinations tend to improve therapeutically relevant selectivity. *Nat. Biotechnol.*, **27**, 659–666.
23. Zhou,X., Seto,S.W., Chang,D., Kiat,H., Razmovski-Naumovski,V., Chan,K. and Bensoussan,A. (2016) Synergistic effects of chinese herbal medicine: a comprehensive review of methodology and current research. *Front. Pharmacol.*, **7**, 201.
24. Liu,J. and Wang,Z. (2015) Diverse array-designed modes of combination therapies in Fangjiomics. *Acta Pharmacol. Sin.*, **36**, 680–688.