https://doi.org/10.1093/nar/gkab1011

HIT 2.0: an enhanced platform for Herbal Ingredients' Targets

Deyu Yan[1,†], Genhui Zheng[1,†], Caicui Wang[1], Zikun Chen[1], Tiantian Mao[1], Jian Gao[2,3], Yu Yan[1], Xiangyi Chen[1], Xuejie Ji[1], Jinyu Yu[1], Saifeng Mo[1], Haonan Wen[1], Wenhao Han[1], Mengdi Zhou[1], Yuan Wang[1], Jun Wang[1], Kailin Tang[1,*] and Zhiwei Cao[4,*]

[1]Dept. of Gastroenterology, Shanghai Tenth People's Hospital, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China, [2]International Human Phenome Institutes (Shanghai), Shanghai, China, [3]Department of Thoracic Surgery, Fudan University Shanghai Cancer Center, Shanghai, China and [4]School of Life Sciences, Fudan University, Shanghai 200092, China

Received September 14, 2021; Revised October 08, 2021; Editorial Decision October 11, 2021; Accepted November 15, 2021

ABSTRACT

Literature-described targets of herbal ingredients have been explored to facilitate the mechanistic study of herbs, as well as the new drug discovery. Though several databases provided similar information, the majority of them are limited to literatures before 2010 and need to be updated urgently. HIT 2.0 was here constructed as the latest curated dataset focusing on Herbal Ingredients' Targets covering PubMed literatures 2000–2020. Currently, HIT 2.0 hosts 10 031 compound-target activity pairs with quality indicators between 2208 targets and 1237 ingredients from more than 1250 reputable herbs. The molecular targets cover those genes/proteins being directly/indirectly activated/inhibited, protein binders, and enzymes substrates or products. Also included are those genes regulated under the treatment of individual ingredient. Crosslinks were made to databases of TTD, DrugBank, KEGG, PDB, UniProt, Pfam, NCBI, TCM-ID and others. More importantly, HIT enables automatic Target-mining and My-target curation from daily released PubMed literatures. Thus, users can retrieve and download the latest abstracts containing potential targets for interested compounds, even for those not yet covered in HIT. Further, users can log into 'My-target' system, to curate personal target-profiling on line based on retrieved abstracts. HIT can be accessible at http://hit2.badd-cao.net.

INTRODUCTION

Being a rich source of drug candidates, herbal active ingredients play a critical role in the development of new drugs. From 1981 to 2019, 33.6% of the drugs approved by the FDA were reported to be derived from natural products or their derivatives (1). To better understand the interaction between herbal compounds and molecular targets, the first herbal-ingredient-target database, HIT, was established in 2010 via manual curation of 1301 literature-described targets for herbal compound from 3250 literatures, with convenient links to therapeutic targets database (TTD) and Drugbank etc (2,3). The target information of HIT has been extensively exploited to study the mechanism of natural compounds, as well as to make discoveries from herbal medicine. For instance, based on HIT, Luo Y et al. revealed the therapeutic mechanism of cryptotanshinone in the treatment of liver cancer (4). And Wang *et al.* identified potential targets for asthma according to the clinical efficacy of TCM formulations (5).

During the past ten years (2011–2021), there was an explosive increase in the studies on the natural ingredients and their targets. A number of databases were then constructed covering herbal compound and target interactions. Notably, a nice exemplary database for natural products, NPASS (6), provides experimentally-determined quantitative activity records for natural products, including nearly 2,000 herbal ingredient-target pairs for about 700 herbal ingredients. Other herbal databases also included important information of herbal ingredient-target pairs (7–11), but most of them downloaded and incorporated early HIT data, appended with predicted targets, such as HERB, TCMID, TCMSP and SymMap (7–10). These databases have greatly enriched the target diversity for herbal ingredients. However, the literature-described targets of herbal ingredients

[*]To whom correspondence should be addressed. Tel: +86 21 65980296; Email: zwcao@fudan.edu.cn
Correspondence may also be addressed to Kailin Tang. Tel: +86 21 65980296; Email: kltang@tongji.edu.cn
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

being carefully curated, though valuable, remain limited and need to be updated urgently and regularly.

Yet those fresh evidences of ingredient-target interactions may be published in everyday-released literatures, while the traditional databases have difficulties in keeping track of the latest results. Users often need to take great efforts manually searching the up-to-date literatures not yet covered by current databases and dig targets out to tie up with database items, which is highly time-consuming. Thus, it is necessary to propose a new platform to provide not only the regularly updated targets, but also the real-time checking for potential target from daily-released papers.

Here, we introduced such a platform of HIT 2.0 for the above purpose. In this version, advanced text mining algorithms and rigorous curation were comprehensively employed. In addition to calibrating HIT 1.0 data from literatures between 2000 and 2010, our curation team made a complete refreshment by adding literatures between 2010 and 2020, resulting in almost twice the data abundancy as before, plus new features of target confidence indicators. More importantly, the text-mining system of compound-target suggestion is now open to users, where researchers can retrieve the most-related literatures via 'Target-mining' for any natural compounds. At last, HIT 2.0 provides an on-line function of 'My-target' for personal curation and downloading.

MATERIALS AND METHODS

Data source

Similar to HIT 1.0, the ingredient information was sourced from the widely used TCM-ID database of the 2020 updated version (12), which covers 2751 herbs and 7375 herb ingredients. Compound aliases were derived from the Chemical Abstracts Service (CAS). The same set of 59 keywords was used as that in HIT 1.0 to describe interactions between compounds and molecular targets (Supplementary Table S1). Some keywords are nouns describing the interaction (Type A), while the others (Type B) are phrases describing the specific effect, such as 'inhibit the activity of' proteins.

Text mining

Text mining of NLP was used to identify the targets of herbal ingredients in the literature, with workflow being illustrated as follows (Figure 1):

Step1: Retrieve abstracts in PubMed containing keywords of the name/alias of herb ingredients.

Step2: Annotate gene/protein entities in the abstract using PubTator Central (13), an automated annotation platform for biomedical entities. These genes/proteins may be targets of ingredients, and only abstracts containing gene/protein entities will be retained.

Step3: Parse the full abstract into sentences. Screen out those sentences based on either of below two rules. Rule1: 'Compound name' AND 'any word in type A' AND 'Gene'. For instance, the sentence 'EGCG is a novel Hsp90 inhibitor'. Rule2: 'Compound name' AND 'any word in type B interaction' AND 'any word in type B effect' AND 'Gene'. For instance, the sentence 'Procyanidin B2 directly inhibited MT1-MMP activity.'

Step4: Use the Stanford Parser engine (14) to extract sentence syntactic structure and grammatical relations (15). Only those sentences are retained where 'compound', 'gene' and 'keyword' are detected in a directional dependency tree path.

Manual curation

Finally, a curation team of 11 Ph.D. candidates reviewed 17 000 sentences, and each item was double checked by at least two candidates. The entries with consistence from two curators were remained. While those of disagreement were reviewed by a third curator, with the final result being decided by a majority vote.

RESULTS

The construction of HIT 2.0 was based on literature mining and manual curation. In the part of literature mining, PubMed abstracts were firstly retrieved containing keywords of the name/alias of herbal ingredients. Then the abstract was parsed into sentences and only those sentences containing herbal ingredients, genes, and keywords were kept for further curation. For convenience, an online platform was set up where each curator has their own account. Any time after log-in, curators can review the PubMed tasks completed, and to-be-completed. Now the whole platform has been opened as 'Target-mining' and 'My-target' system, enabling users to identify those most related literatures and keep close tracking of the latest targets for their interested compounds.

Data updating

In this version, more advanced text-mining algorithms of Natural language processing (NLP) and rigorous manual curation were applied using the same set of keywords as that of HIT 1.0 (16). NLP has been widely applied to biomedical text mining. Dependent syntactic analysis enables sentence structure parsing to highlight the relationships between medical entities (15). The NLP algorithm is used in HIT 2.0 to determine whether compounds may interact with genes/proteins. Initially, 7100 abstracts were obtained from PubMed after text-mining system. After the online curation, a total of 10 031 compound-target pairs were produced, involving 2208 molecular targets and 1,237 compounds from 1250 herbs. Interestingly, 56 miRNA genes have been included into our targeting list, mainly with modes of up/down regulated genes.

The types of compound-target interactions cover 10 categories: indirectly inhibit/activate, up/down-regulated gene, directly inhibit/activate, binders, enzyme substrates, enzyme products and other. Quality indicators of compound-target interactions have been developed covering the nature of the interaction, the number of literatures supporting the same pair, and the citing reports of each literature. Based on above, users can choose those preferred for further analysis. Each compound-target pair can be viewed with a key description parsed from the sourcing literature. Compounds are linked to PubChem (17) and ChEMBL (18) databases. As for targets, Crosslinks have been made to databases of

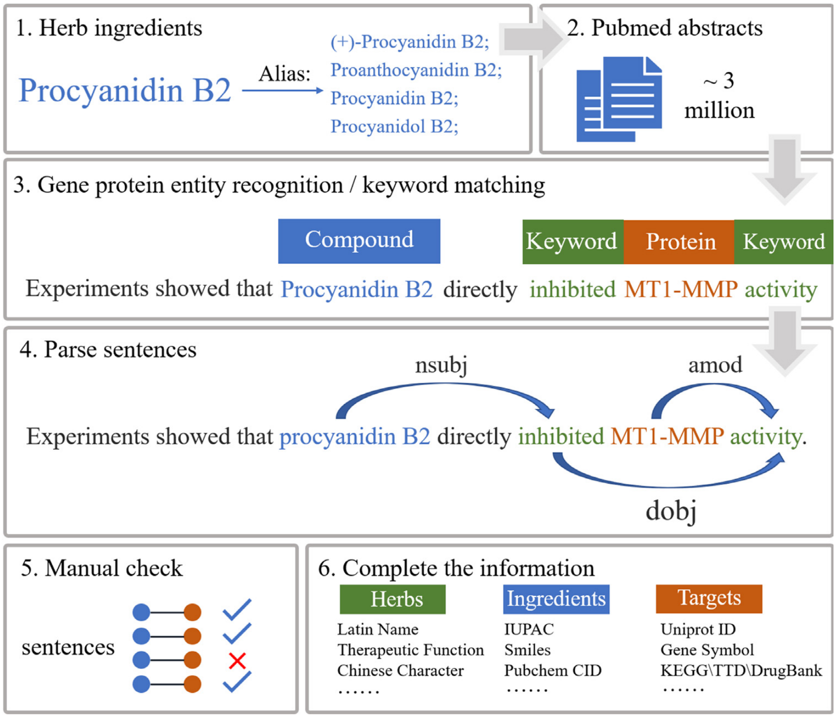


Figure 1. Workflow of HIT 2.0. (1, 2) Retrieve PubMed using different names of herbal ingredients. (3) Mine PubMed abstracts to identify gene/protein entities. (4) Detect whether 'compound', 'gene' and 'keyword' are in a directional dependency tree path. 5&6. Manual check and complete the information.

TTD (2), DrugBank (3), KEGG (19), PDB (20), UniProt (21), Pfam (22), NCBI, TCM-ID (12) and others for more detailed information.

Compared to similar herb databases containing literature-described targets, HIT 2.0 has doubled the data abundancy of HIT 1.0, forming a nice complement to the previous databases. Currently, it is also the largest database in terms of curated ingredient-target pairs for herbs, as Table 1 shows.

HIT 2.0 allows keyword search and compound similarity search. The searching interface and results pages are illustrated in Figure 2. Keyword search is available for herb information [Chinese pinyin, Chinese characters, Latin name and English name], ingredients information [different names, CAS number and CID number] and target information [gene/protein name, gene symbol, UniProt ID]. Auto-completion and fuzzy search are supported for keyword search. Besides that, similarity search can also be made via compound structures, with a Tanimoto coefficient above 70% as a cut-off. Both the SMILES formula and the artificially drawn structure by build-in software Ketcher (https://lifescience.opensource.epam.com/ketcher/) can be used as input.

Target-mining and My-target curation

As PubMed literatures are released every day, there is a constant need for researchers to check the latest evidences after 2020 for their interested compound. So far, no software becomes available to mine the literatures containing potential target-ingredient interactions. For the convenience of curation, two options have been provided for users.

Target-mining function was built-in enabling users to retrieve related PubMed abstracts published from 2010 till the daily updates for any compounds (Figure 3A and B). It was realized that the abstracts identified by text-mining may contain false positives. Thus HIT 2.0 provides on-line My-target curation function to further check the detailed sentences in sourcing abstracts (Figure 3C). The sentences containing interesting entities have been highlighted in different colors, so that key items can be easily spotted together with the interaction types. A link to PubMed with full abstract is also provided.

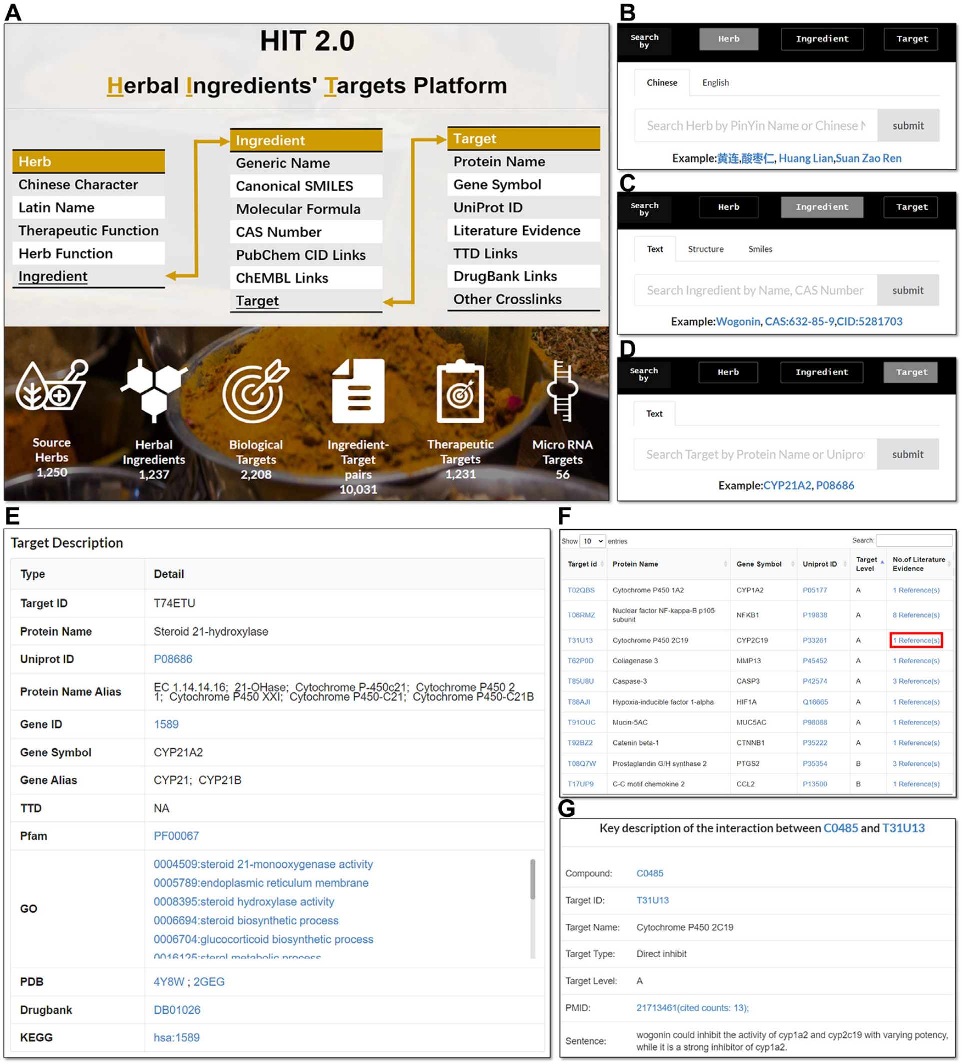
In brief, 'Target-mining' was designed to efficiently and precisely retrieve those most-related abstracts based on our empirically derived rules combined with advanced text-mining algorithm. Via this, users can retrieve and download the latest literatures for local checking. Alternatively, users can choose 'My-target curation' to make curation on-line. Each user has their own account. Anytime after log-in, users can review the tasks completed and continue to finish the full task.

DISCUSSION

Complement to other drug and drug-target databases, herbal databases containing literature-described targets have served as the primary source for mechanistic study of natural products by providing rich sets of information between medicinal herbs, active compounds and molecular targets under different experimental conditions. With a steady literature accumulation describing new evidences in the past decade, however, the previous versions need to make timely updating to meet up with the pressing needs. In

Table 1. Overview of the literature-described targets from peering databases

	Published year	Literature-described targets	Herbal ingredients	Herbal ingredient-target activity pairs	Sourcing literatures
HIT 2.0		2208	1237	10 031	7100 PubMed abstracts
HIT	2011	1301	586	5208	3250 PubMed abstracts
HERB	2020	1241	370	4815	1966 PubMed abstracts
TCMID	2013	680	/	/	4500 Chinese Literatures
NPASS	2017	464	719	1936	1288 PubMed abstracts
TCMSP	2014	*3311 predicted bySysDT*	*29 384*	*84 260 predicted bySysDT*	/



Target Description

Type	Detail
Target ID	T74ETU
Protein Name	Steroid 21-hydroxylase
Uniprot ID	P08686
Protein Name Alias	EC 1.14.14.16; 21-OHase; Cytochrome P-450c21; Cytochrome P450 2 1; Cytochrome P450 XXI; Cytochrome P450-C21; Cytochrome P450-C21B
Gene ID	1589
Gene Symbol	CYP21A2
Gene Alias	CYP21; CYP21B
TTD	NA
Pfam	PF00067
GO	0004509:steroid 21-monooxygenase activity 0005789:endoplasmic reticulum membrane 0008395:steroid hydroxylase activity 0006694:steroid biosynthetic process 0006704:glucocorticoid biosynthetic process 0016125:sterol metabolic process
PDB	4Y8W ; 2GEG
Drugbank	DB01026
KEGG	hsa:1589

Show 10 entries Search:

Target id	Protein Name	Gene Symbol	Uniprot ID	Target Level	No.of Literature Evidence
T02QBS	Cytochrome P450 1A2	CYP1A2	P05177	A	1 Reference(s)
T06RMZ	Nuclear factor NF-kappa-B p105 subunit	NFKB1	P19838	A	8 Reference(s)
T31U13	Cytochrome P450 2C19	CYP2C19	P33261	A	1 Reference(s)
T62P0D	Collagenase 3	MMP13	P45452	A	1 Reference(s)
T85U8U	Caspase-3	CASP3	P42574	A	3 Reference(s)
T88AJI	Hypoxia-inducible factor 1-alpha	HIF1A	Q16665	A	1 Reference(s)
T91OUC	Mucin-5AC	MUC5AC	P98088	A	1 Reference(s)
T92BZ2	Catenin beta-1	CTNNB1	P35222	A	1 Reference(s)
T08Q7W	Prostaglandin G/H synthase 2	PTGS2	P35354	B	3 Reference(s)
T17UP9	C-C motif chemokine 2	CCL2	P13500	B	1 Reference(s)

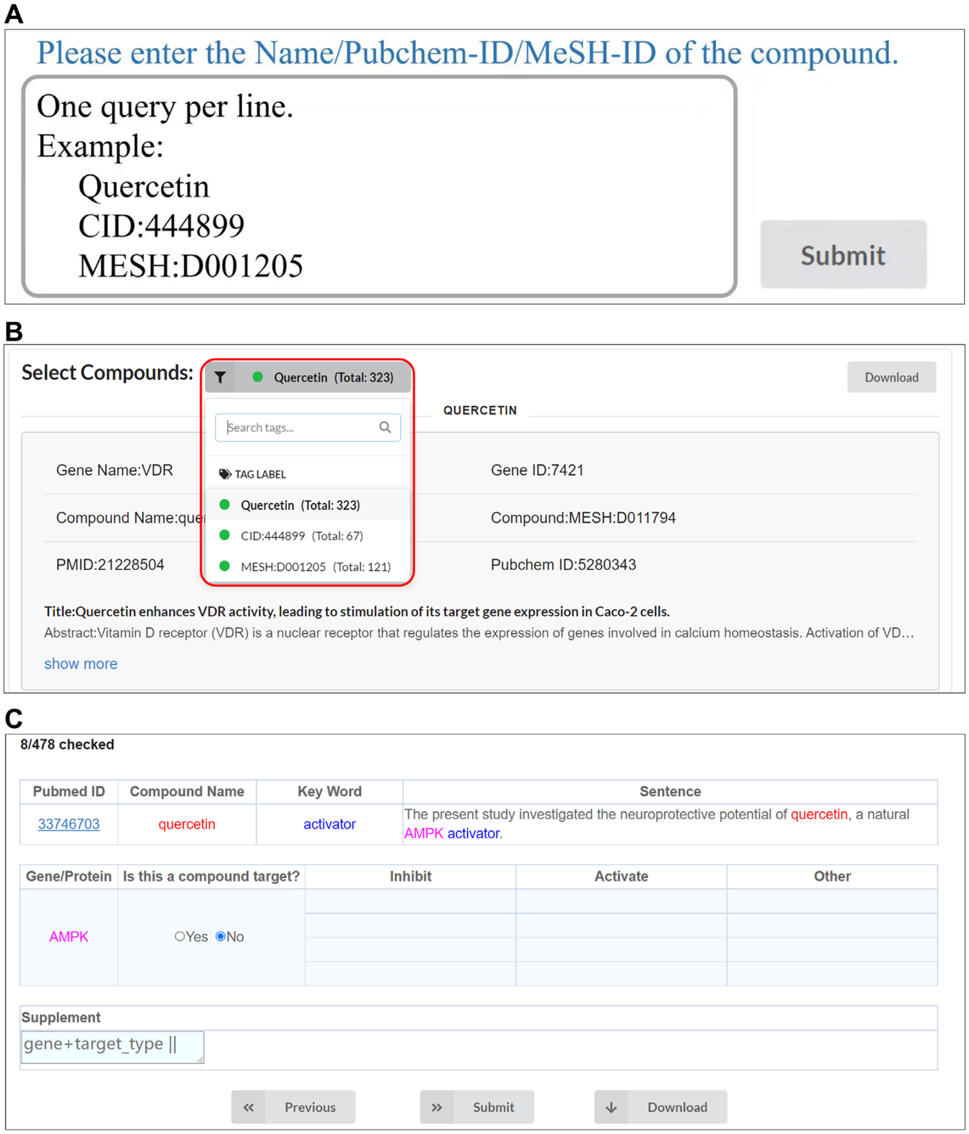
Key description of the interaction between C0485 and T31U13

Compound:	C0485
Target ID:	T31U13
Target Name:	Cytochrome P450 2C19
Target Type:	Direct inhibit
Target Level:	A
PMID:	21713461(cited counts: 13);
Sentence:	wogonin could inhibit the activity of cyp1a2 and cyp2c19 with varying potency, while it is a strong inhibitor of cyp1a2.

Figure 2. Searching and resulting pages in HIT 2.0. (**A**) Database structure and data statistics. (**B**) Herbs can be searched via keywords such as Chinese Pinyin, Chinese characters and Latin names. (**C**) Herbal ingredients can be searched via structure similarity or keywords of name, CID and CAS number. (**D**) Targets can be searched via keywords of gene/protein name, gene symbol and Uniprot ID. (**E**) Detailed information of the targets. (**F**) Additional targets of the compound. (**G**) ‘ Literature evidence ’ provides the key descriptions parsed from sourcing literatures.

this paper, HIT 2.0 was thus constructed to maintain a regular updating by adding another ten years, marking as the latest and the largest one regarding curated ingredient-target pairs for herbs. Meanwhile, we set up Target-mining and My-target curation system based on technology of natural processing language, allowing researchers to keep tracking the latest evidences and curate personal targets of interested compounds at convenient time. The launch of HIT 2.0 will be an important addition to bridge herbs ingredients and FDA approved drugs via molecular targets and may facilitate the discovery of new druggable molecules, as well as to identify potential therapeutic targets.

One key technology applied in HIT 2.0 was the auto text-mining tools. Currently, there are several annotating



Please enter the Name/Pubchem-ID/MeSH-ID of the compound.

One query per line.
Example:
Quercetin
CID:444899
MESH:D001205

Submit

Select Compounds: Quercetin (Total: 323)

Download

Search tags...

TAG LABEL

- ● Quercetin (Total: 323)
- ● CID:444899 (Total: 67)
- ● MESH:D001205 (Total: 121)

QUERCETIN

Gene Name:VDR	Gene ID:7421
Compound Name:que	Compound:MESH:D011794
PMID:21228504	Pubchem ID:5280343

Title:Quercetin enhances VDR activity, leading to stimulation of its target gene expression in Caco-2 cells.

Abstract:Vitamin D receptor (VDR) is a nuclear receptor that regulates the expression of genes involved in calcium homeostasis. Activation of VD...

show more

8/478 checked

Pubmed ID	Compound Name	Key Word	Sentence
33746703	quercetin	activator	The present study investigated the neuroprotective potential of quercetin, a natural AMPK activator.

Gene/Protein	Is this a compound target?	Inhibit	Activate	Other
AMPK	○Yes ●No			

Supplement

gene+target_type ||

« Previous » Submit ↓ Download

Figure 3. Target-mining and My-target curation system. (**A**) The interface of Target-mining function. Compound name, MeSH ID and Pubchem ID can be submitted to retrieve potential targets. (**B**) PubMed abstract retrieved by Target-mining. (**C**) The interface of My-target curation system.

tools to recognize the biomedical entities, such as PubTator Central (13), HunFlair (23), and ScispaCy (24). Among them, PubTator Central was developed by NCBI to annotate PubMed abstracts automatically and allows users to download annotations in bulk via PubMed ID lists. In addition, the annotated genes have direct links to the Gene database, which allows HIT 2.0 to access more details of the targets without being installed. Considering the overall convenience, simplicity and complexity, PubTator central was chosen for HIT 2.0.

Meanwhile, the last decade has witnessed an extensive application of molecular targets into functional explanation for herbal compounds. Though the reliability of literature-evidenced targets may be more accurate than computationally-predicted targets, it is aware that the results collected into HIT are mostly from *in vitro* rather than *in vivo* experiments. In fact, these natural compounds are likely metabolized into different forms which may change the targeting profiles (25). Furthermore, the biological activity of an ingredient is often related to not only the group of molecular targets, but also the network interactions among targets. In this sense, HIT 2.0 may serve as a valuable start in deriving collective functions of herbal components, particularly to herbs and herbal formulas. Towards this direction, HIT will continue to enrich the data abundancy and subsequent analysis to maintain high-quality resources for domain research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Key R&D Program of China [2017YFC1700200, 2019YFA0905900]; National Natural Science Foundation of China [32070657, 81830080]. Funding for open access charge: National Key R&D Program of China [2017YFC1700200].

Conflict of interest statement. None declared.

REFERENCES

1. Newman,D.J. and Cragg,G.M. (2020) Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.*, **83**, 770–803.
2. Wang,Y., Zhang,S., Li,F., Zhou,Y., Zhang,Y., Wang,Z., Zhang,R., Zhu,J., Ren,Y., Tan,Y. *et al.* (2020) Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res*, **48**, D1031–D1041.
3. 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.
4. Luo,Y., Song,L., Wang,X., Huang,Y., Liu,Y., Wang,Q., Hong,M. and Yuan,Z. (2020) Uncovering the mechanisms of cryptotanshinone as a therapeutic agent against hepatocellular carcinoma. *Front. Pharmacol.*, **11**, 1264.
5. Wang,Y., Chen,Y.J., Xiang,C., Jiang,G.W., Xu,Y.D., Yin,L.M., Zhou,D.D., Liu,Y.Y. and Yang,Y.Q. (2020) Discovery of potential asthma targets based on the clinical efficacy of Traditional Chinese Medicine formulas. *J. Ethnopharmacol.*, **252**, 112635.
6. Zeng,X., Zhang,P., He,W., Qin,C., Chen,S., Tao,L., Wang,Y., Tan,Y., Gao,D., Wang,B. *et al.* (2018) NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res.*, **46**, D1217–D1222.
7. Fang,S., Dong,L., Liu,L., Guo,J., Zhao,L., Zhang,J., Bu,D., Liu,X., Huo,P., Cao,W. *et al.* (2021) HERB: a high-throughput experiment- and reference-guided database of traditional Chinese medicine. *Nucleic Acids Res.*, **49**, D1197–D1206.
8. 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.
9. 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.
10. Wu,Y., Zhang,F., Yang,K., Fang,S., Bu,D., Li,H., Sun,L., Hu,H., Gao,K., Wang,W. *et al.* (2019) SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping. *Nucleic Acids Res.*, **47**, D1110–D1117.
11. Zhang,R.Z., Yu,S.J., Bai,H. and Ning,K. (2017) TCM-Mesh: The database and analytical system for network pharmacology analysis for TCM preparations. *Sci. Rep.*, **7**, 2821.
12. 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.
13. Wei,C.H., Allot,A., Leaman,R. and Lu,Z. (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.*, **47**, W587–W593.
14. Marneffe,M.-C., MacCartney,B. and Manning,C. (2006) In: *Generating Typed Dependency Parses from Phrase Structure Parses*. http://www.lrec-conf.org/proceedings/lrec2006/pdf/440_pdf.pdf.
15. Kveler,K., Starosvetsky,E., Ziv-Kenet,A., Kalugny,Y., Gorelik,Y., Shalev-Malul,G., Aizenbud-Reshef,N., Dubovik,T., Briller,M., Campbell,J. *et al.* (2018) Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed. *Nat. Biotechnol.*, **36**, 651–659.
16. 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 ingredients to targets. *Nucleic Acids Res.*, **39**, D1055–D1059.
17. Kim,S., Chen,J., Cheng,T., Gindulyte,A., He,J., He,S., Li,Q., Shoemaker,B.A., Thiessen,P.A., Yu,B. *et al.* (2021) PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.*, **49**, D1388–D1395.
18. Mendez,D., Gaulton,A., Bento,A.P., Chambers,J., De Veij,M., Felix,E., Magarinos,M.P., Mosquera,J.F., Mutowo,P., Nowotka,M. *et al.* (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.*, **47**, D930–D940.
19. Kanehisa,M., Furumichi,M., Sato,Y., Ishiguro-Watanabe,M. and Tanabe,M. (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.
20. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chen,L., Crichlow,G.V., Christie,C.H., Dalenberg,K., Di Costanzo,L., Duarte,J.M. *et al.* (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, **49**, D437–D451.
21. UniProt,C. (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
22. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J. *et al.* (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res*, **49**, D412–D419.
23. Weber,L., Sanger,M., Munchmeyer,J., Habibi,M., Leser,U. and Akbik,A. (2021) HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*,**37**, 2792–2794.
24. Neumann,M., King,D., Beltagy,I. and Ammar,W. (2019) In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. https://aclanthology.org/W19-50.
25. Kang,H., Tang,K., Liu,Q., Sun,Y., Huang,Q., Zhu,R., Gao,J., Zhang,D., Huang,C. and Cao,Z. (2013) HIM-herbal ingredients in-vivo metabolism database. *J. Cheminform.*, **5**, 28.