SCIENTIFIC REPORTS

OPEN

In silico identification of anti-cancer compounds and plants from traditional Chinese medicine database

Received: 07 December 2015
Accepted: 18 April 2016
Published: 05 May 2016

Shao-Xing Dai[1,2], Wen-Xing Li[1,3], Fei-Fei Han[1,2], Yi-Cheng Guo[1,4], Jun-Juan Zheng[1,2], Jia-Qian Liu[1,2], Qian Wang[1,2], Yue-Dong Gao[5], Gong-Hua Li[1,2] & Jing-Fei Huang[1,2,6,7]

There is a constant demand to develop new, effective, and affordable anti-cancer drugs. The traditional Chinese medicine (TCM) is a valuable and alternative resource for identifying novel anti-cancer agents. In this study, we aim to identify the anti-cancer compounds and plants from the TCM database by using cheminformatics. We first predicted 5278 anti-cancer compounds from TCM database. The top 346 compounds were highly potent active in the 60 cell lines test. Similarity analysis revealed that 75% of the 5278 compounds are highly similar to the approved anti-cancer drugs. Based on the predicted anti-cancer compounds, we identified 57 anti-cancer plants by activity enrichment. The identified plants are widely distributed in 46 genera and 28 families, which broadens the scope of the anti-cancer drug screening. Finally, we constructed a network of predicted anti-cancer plants and approved drugs based on the above results. The network highlighted the supportive role of the predicted plant in the development of anti-cancer drug and suggested different molecular anti-cancer mechanisms of the plants. Our study suggests that the predicted compounds and plants from TCM database offer an attractive starting point and a broader scope to mine for potential anti-cancer agents.

Cancer, also known as a malignant tumor, is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body. The hallmarks of cancer comprise six biological capabilities to support the development of human tumors, which include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis[1,2]. Cancer is one of the major causes of death worldwide where the number of cancer patient is in continuous rise. There are over 100 different known cancers that affect humans, and each is classified by the type of cell that is initially affected[3]. In 2012 about 14.1 million new cases of cancer occurred globally (not including skin cancer other than melanoma). It caused about 8.2 million deaths or 14.6% of all human deaths[4]. By 2030, it is predicted that there will be 26 million new cancer cases and 17 million cancer deaths per year[5].

Today, despite considerable efforts, cancer still remains an aggressive killer worldwide. The most common and highly effective methods of cancer treatment are surgery, chemotherapy and radiotherapy[6]. However, these therapies have numerous limitations and drawbacks[7]. Most cancer patients are diagnosed too late to undergo surgery because of poor diagnosis and other factors. Chemotherapy and radiotherapy have serious side effects and complications such as fatigue, pain, diarrhea, nausea, vomiting, and hair loss[7]. Furthermore, chemotherapy and radiotherapy can result in gradual resistance of cancer cells against treatment[8].

[1]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, Yunnan, China. [2]Kunming College of Life Science, University of Chinese Academy of Sciences, Beijing 100049, China. [3]Institute of Health Sciences, Anhui University, Hefei 230601, Anhui, China. [4]School of Life Sciences, University of Science and Technology of China, Hefei, Anhui 230027, China. [5]Kunming Biological Diversity Regional Center of Instruments, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China. [6]KIZ-SU Joint Laboratory of Animal Models and Drug Development, College of Pharmaceutical Sciences, Soochow University, Kunming 650223, Yunnan, China. [7]Collaborative Innovation Center for Natural Products and Biological Drugs of Yunnan, Kunming 650223, Yunnan, China. Correspondence and requests for materials should be addressed to G.-H.L. (ligonghua@mail.kiz.ac.cn) or J.-F.H. (huangjf@mail.kiz.ac.cn)

Therefore there is a constant demand to develop new, effective, and affordable anti-cancer drugs[9]. Medicinal plants constitute a common alternative for cancer treatment in many countries around the world[10–13]. There are more than 2000 plants used in the traditional Chinese medicine (TCM) according to the TCM database@taiwan (http://tcm.cmu.edu.tw/)[14]. These medicinal plants were used for treatment of various diseases include cancer for thousand years in China[15–19]. Many TCM-derived anti-cancer products have been used in western medicine[20–28]. These include vinblastine, vincristine, paclitaxel, camptothecin, epipodophyllotoxin and so on. Vinblastine and vincristine, as the bisindole alkaloids isolated from *Catharanthus roseus*, are the first agents to advance into clinical use for treatment of spleen cancer, liver cancer and childhood leukemia. Paclitaxel, originally isolated from the bark of *Taxus brevifolia*, has also been found in *Taxus chinensis*. It was launched in 1992 and was the best-selling anti-cancer drug in the USA in 2002[8]. Another important class of anti-cancer drugs (topotecan, irinotecan, belotecan, 9-Nitrocamptothecin, and gimatecan) are derived from camptothecin which was isolated from the Chinese ornamental tree *Camptotheca acuminate*[8,29]. Epipodophyllotoxin is also an important class of natural product for development of anti-cancer drugs. Etoposide, teniposide and etopophos are semi-synthetic derivatives of epipodophyllotoxin[8]. They are approved for treatment of choriocarcinoma, lung cancer, ovarian and testicular cancers, lymphoma, acute myeloid leukemia, and bladder cancer[6].

TCM is undoubtedly a valuable resource for identifying novel anti-cancer agents[30]. Regrettably, only a small portion of medicinal plants in the TCM database has been fully phytochemically investigated. It is interest to systematic explore and evaluate the anti-cancer potential of all the plants in the TCM database. However, it is a tedious, expensive and time-consuming process because that it involves screening of large molecular library by experiment. Therefore, the time and money-saving way is that the plants in the TCM database are firstly filtered by the computational analysis of the anti-cancer potential, then evaluated by experiment. The aim of the current investigation is to analyze the anti-cancer potential of all the plants in the TCM database by using cheminformatics, and then identify the anti-cancer compounds and plants from the TCM database in silico. We started with the TCM Database@Taiwan, which is currently the world's largest non-commercial TCM database[14]. The database contains the relationship between more than 20,000 pure compounds and more than 2000 plants. We first predicted anti-cancer compounds in the database by using our previously published method termed Cancer Drug (CDRUG)[31]. We then determined the anti-cancer plants by performing the anti-cancer activity enrichment analysis (ACEA)[32]. Each of the anti-cancer plants was significantly enriched with anti-cancer compounds. Thus, the identified anti-cancer plants provide important clues and direction for the development of anti-cancer drugs.

Results

Prediction of anti-cancer compounds from TCM Database@Taiwan. A total of 21334 compounds from 2402 plants were downloaded from TCM Database@Taiwan. The anti-cancer activity of these compounds was predicted using CDRUG. Finally, a total of 5278 compounds were predicted as anti-cancer compounds ($P < 0.05$), which is accounting for 25% (5278/21334) of all compounds in the database. Further careful observation, we found the top 346 compounds were identical to those compounds which have been proven active in the 60 cell lines test reported by NCI-60 DTP project[33]. Most of the top 346 compounds have the inhibition rate of growth $>50\%$ at less than the dose of 10^{-5} mol/L. The mean logGI50 value (the 50% growth inhibition concentration) of the top 346 compounds is -5.73 with standard deviation 0.89. Among the top 346 compounds, two compounds paclitaxel and homoharringtonine have already been approved for the treatment of various cancers. The logGI50 values of drugs paclitaxel and homoharringtonine are -7.74 and -7.152, respectively.

Similarity of the predicted anti-cancer compounds with the anti-cancer drugs. Since the compounds identified above were predicted to have anti-cancer activity, we performed a systematic analysis of the similarity between these compounds and the anti-cancer drugs in preclinical, clinical and approved stages from the database of Thomson Reuters Integrity. We got 127, 425 and 219 anti-cancer drugs in preclinical, clinical and approved stages, respectively (**Dataset1** Table S2). Then the similarities of the 5278 compounds against all the anti-cancer drugs of the three types were calculated (see **Methods**). Two compounds are considered structurally similar if their fingerprints have a Tc of 0.70 or greater. We found that 4025 (76%) of the 5278 compounds have similarity (Tc 0.70, MACCS fingerprint) with the anti-cancer drugs in preclinical stage. Similarly, 4406 (83%) and 3952 (75%) of the 5278 compounds have similarity with the anti-cancer drugs in clinical and approved stages, respectively. These results demonstrate the power of CDRUG for prediction of anti-cancer compound. It also shows the importance of these plant-derived compounds in the development of anti-cancer drugs.

Structural characteristics of the predicted active compounds. Orally administered drugs are more likely in areas of chemical space defined by a limited range of molecular properties which were encapsulated in Lipinski's 'rule of five'[34]. Lipinski's rule states that, historically, 90% of orally absorbed drugs had fewer than 5 H-bond donors, less than 10 H-bond acceptors, molecular weight of less than 500 daltons and AlogP values of less than 5. To compare the predicted active compounds with cancer drugs, the four properties and other important properties (number of rotatable bonds, rings, aromatic rings) were calculated in our study (Fig. 1). The distributions of AlogP and molecular weight for the two classes of compounds are highly similar and overlapped (Fig. 1A). In total, 73% of the predicted active compounds have AlogP less than 5 compared with 85% for cancer drugs. In contrast, only 50% and 57% of molecules have a molecular weight less than 500 daltons for the predicted active compounds and cancer drugs, respectively. It suggests the molecules with a molecular weight of more than 500 daltons are also suitable to develop anti-cancer drugs. The major differences between the two classes of compounds emerge when the number of rings and aromatic rings is considered (Fig. 1B,C). 40% of the predicted active compounds have five or more rings compared with 18% for the cancer drugs. Conversely, only 6% of the predicted active compounds have two or more aromatic rings compared with 40% for the cancer drugs. The ratios of the number of rings and aromatic rings are 8.39:1 and 1.67:1 for the predicted active compounds

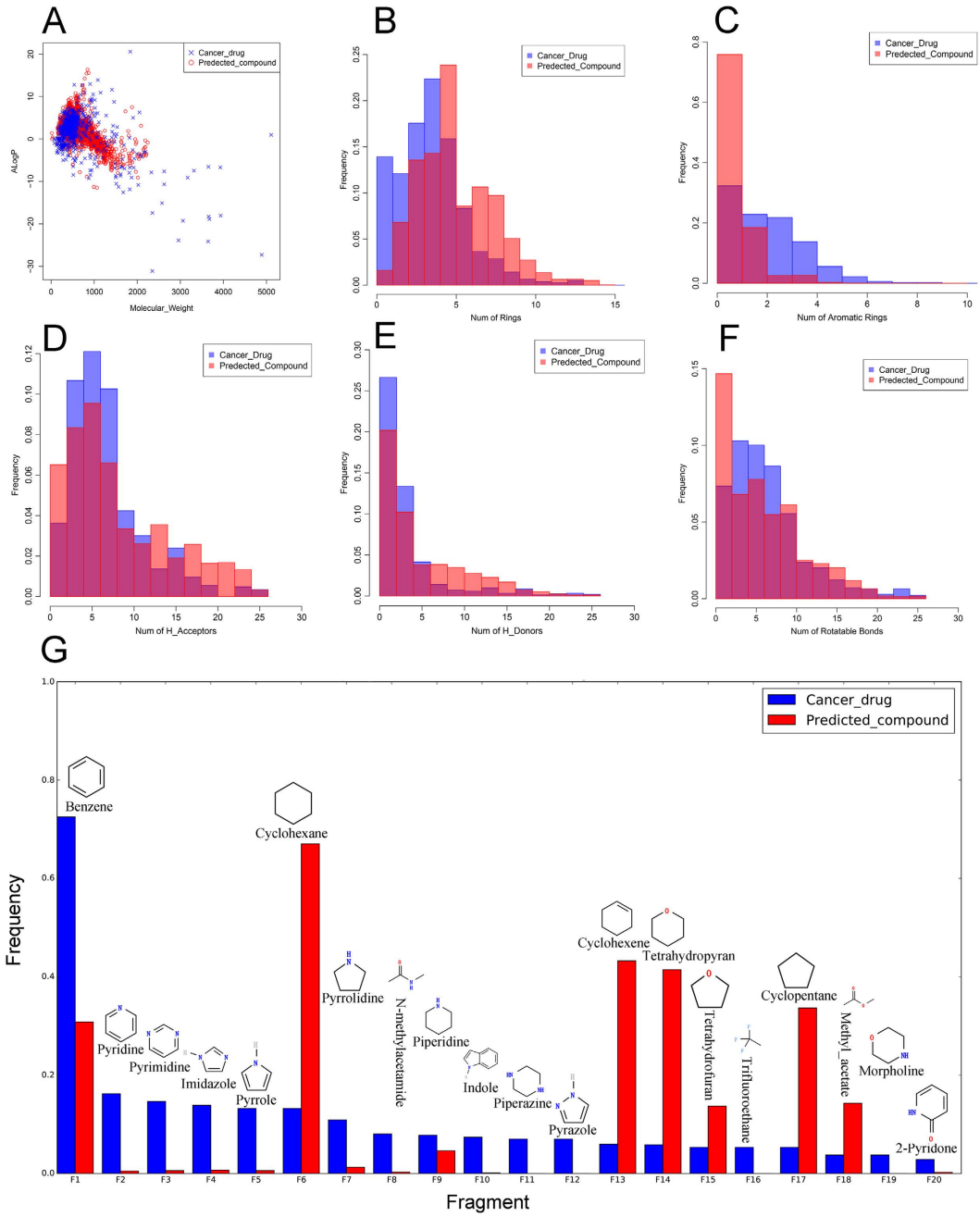

Figure 1. Structural characteristics of the predicted active compounds and cancer drugs. (**A**) Scatter plot of molecular weight against ALogP. (**B**) Histogram plot of the number of rings. (**C**) Histogram plot of the number of aromatic rings. (**D**) Histogram plot of the number of H-bond acceptors. (**E**) Histogram plot of the number of H-bonds donors. (**F**) Histogram plot of the number of rotatable bonds. (**G**) The bar plot of the top 20 common fragments and their frequency (F1-F20). In all plot, the cancer drugs and predicted active compounds were colored by blue and red, respectively.

and cancer drugs, respectively. The predicted active compounds tend toward a high ratio of the number of rings and aromatic rings compared with the cancer drugs. The distributions of the other three molecular properties (number of H-bond donors, H-bond acceptors and rotatable bonds) are similar between the two classes of compounds (Fig. 1D–F).

To further compare the two classes of compounds, the most common fragments and their frequency for these molecules were analyzed. The top 20 common fragments in the cancer drugs were shown in the Fig. 1G. The frequency of these fragments is very different between the two classes of compounds. The frequency of most fragments in the predicted active compounds is less than that in the cancer drugs. For example, the frequency of pyridine, pyrimidine, imidazole, pyrrole and pyrrolidine in the predicted active compounds is extremely low. It is noteworthy that the fragments piperazine, pyrazole, trifluoroethane and morpholine are even absent in the

predicted active compounds. Only six fragments cyclohexane, cyclohexene, tetrahydropyran, tetrahydrofuran, cyclopentane and methyl acetate have higher frequency in the predicted active compounds. The analysis of molecular properties above suggested the predicted active compounds tended toward a high ratio of rings and aromatic rings. This tendency also emerges in the fragments analysis. 73% of the cancer drugs have unsaturated rings benzene. In contrast, 67% of the predicted active compounds have saturated ring cyclohexane. The number of unsaturated rings in the predicted active compounds is far less than that in the cancer drugs. And the number of saturated rings in the predicted active compounds is far more than that in the cancer drugs.

Identification of anti-cancer plants. We have predicted thousands of compounds with anti-cancer activity above. It is worth to identify the plant which is enriched with anti-cancer compounds. The identification of anti-cancer plants is of great value in the introduction, utilization and protection of medicinal plants. It is also important in the development of anti-cancer drugs. Therefore, based on the predicted anti-cancer compounds, we identified 57 anti-cancer plants ($P\_adj < 0.05$) (Table 1) using the method named ACEA. These plants belong to 46 genera and 28 families. Detailed information concerning the anti-cancer plants can be found in Supplementary **Dataset1** Table S3. When checked the family distribution of these plants, we have noticed that the anti-cancer plants were more frequent from the families Araliaceae, Asteraceae, Boraginaceae, Ranunculaceae and Rosaceae. For example, there are 8 anti-cancer plants belonged to family Araliaceae. They are *Panax bipinnatifidum Seem.*, *Panax japonicus*, *Panax notoginseng*, *Panax quinquefolium L.*, *Panax ginseng*, *Aralia elata*, *Oplopanax elatus Nakai*, *Aralia taibaiensis*. These plants have potential ability to kill cancer cells due to the enrichment of anti-cancer compounds. To verify this result, we performed literature survey using Thomson Reuters Web of Science database. We found that many of these plants have been reported to have anti-cancer activity in several studies, such as *Salvia miltiorrhiza*, *Paris polyphylla, Gynostemma pentaphyllum*, *Panax ginseng*, *Panax notoginseng*, *Brucea javanica*, *Platycodon grandiflorum*. Of these plants, *Salvia miltiorrhiza* is the most studied plant for cancer treatment. There are 84 predicted anti-cancer compounds derived from *Salvia miltiorrhiza*. These compounds showed potent activities against various types of cancer including esophageal cancer, gastric cancer, colon cancer, liver cancer, prostate cancer and breast cancer[35–39]. Another more studied plant is *Paris polyphylla Smith* which contains 13 predicted anti-cancer compounds. *Paris polyphylla Smith* has been studied for the treatment of breast cancer, gastric cancer and lung cancer[40–43]. Notably, there are 24 identified anti-cancer plants which were little studied before. These new identified anti-cancer plants are worthy of further studies and provide more chances for the development of cancer drug.

Network of predicted anti-cancer plants and anti-cancer drugs. To show how extend the predicted anti-cancer plants to support the development of anti-cancer drugs, we constructed a network of predicted anti-cancer plants and anti-cancer drugs based on the results above using Cytoscape v3.2. The network connects plant and drug if the compounds in this plant show similarity with this drug (Tc 0.70, MACCS fingerprint). It generated a network which contains 57 plants and 67 anti-cancer drugs (Fig. 2). This network highlights the supportive role of these plants in the development of cancer drugs. All the predicted anti-cancer plants associate with the development of cancer drugs. Some of them appear to be more important and closely related to the development of anti-cancer drugs, such as *Salvia miltiorrhiza*, *Panax ginseng C. A. Mey*, *Brucea javanica*, and *Achyranthes bidentata*. *Salvia miltiorrhiza* connected 6 approved drugs, 10 clinical drugs and 8 preclinical drugs. The six approved drugs are 4-Hydroxyandrostenedione, prednisolone, 17-Methyltestosterone, megestrol acetate, methylprednisolone sodium succinate and bexarotene. These drugs have been used for treatment of breast cancer, lymphoma. Bexarotene is being developed in clinical phase II for treating non-small cell lung cancer. *Panax ginseng C. A. Mey* connected 6 approved drugs, 9 clinical drugs and 6 preclinical drugs. One of the clinical drugs, clinical35 is identical to Ginsenoside K ($TC = 1$) which exist in *Panax ginseng C. A. Mey*. Ginsenoside K is a steroidal saponin in phase I clinical studies at IL-HWA for the treatment of cancer. Similarly, *Brucea javanica* connected 5 approved drugs, 4 clinical drugs and 6 preclinical drugs. *Achyranthes bidentata* connected 4 approved drugs, 6 clinical drugs and 5 preclinical drugs.

Surprisingly, two isolated sub-networks were found in the overall network. The two sub-networks are involved in different drugs, thus maybe different molecular mechanism of anti-cancer. The smaller sub-network contains three plants (*Corydalis incisa*, *Amaryllis belladonna*, and *Thalictrum minus L*) and two approved drugs (approved144: homoharringtonine and approved149: bosutinib). Homoharringtonine was originally isolated from Chinese tree *Cephalotaxus harringtonia*[44]. The three plants and *Cephalotaxus harringtonia* are distributed in different family and order. The diversity of plants and compounds suggests the three plants may provide an alternative resource for discovery of new compounds with activity similar to homoharringtonine. Further studies should be performed to screen the three plants.

Discussion

With the aim of systematic explore and evaluate the anti-cancer potential of all the plants in the TCM database, we identified 5278 anti-cancer compounds in this study. The predicted anti-cancer compounds account for 25% (5278/21334) of all compounds in the database. After calculating similarity, 3952 (75%) of the 5278 compounds have similarity with the approved anti-cancer drugs (Tc 0.70, MACCS fingerprint). It suggests the great value of these predicted anti-cancer compounds. Some new similar drugs may be discovered from these compounds. As natural products, these compounds show less side effects compared with synthetic compound. These compounds can be a ready and effective anti-cancer molecular library. Further experiments should design to screen the library to found the drugs with more active but less side effects.

The compounds which have similarity with the approved anti-cancer drugs can be used to develop me-too drugs. And its opposite, the innovative drugs are developed by using structurally dissimilar compounds and different molecular mechanism. There are about 25% of the 5278 compounds have no similarity with all the

Plant_name	Family	compound	P_adj	literature
Gynostemma pentaphyllum	Cucurbitaceae	36	3.91E-48	39
Platycodon grandiflorum	Campanulaceae	11	2.69E-28	15
Panax japonicus C. A. Mey.	Araliaceae	8	3.65E-26	3
Panax bipinnatifidum Seem.	Araliaceae	42	3.65E-26	0
Panax notoginseng	Araliaceae	14	4.33E-25	28
Annona muricata L.	Annonaceae	39	4.41E-19	10
Pulsatilla chinensis	Ranunculaceae	12	1.70E-14	2
Salvia miltiorrhiza	Lamiaceae	8	6.12E-13	63
Panax quinquefolium L.	Araliaceae	13	1.72E-12	4
Prunella vulgaris	Lamiaceae	13	5.62E-12	13
Polygonatum kingianum	Asparagaceae	50	2.26E-11	0
Patrinia scabiosaefolia	Caprifoliaceae	50	4.69E-11	0
Campsis grandiflora	Bignoniaceae	94	1.05E-10	0
Albizia julibrissin	Fabaceae	32	2.75E-10	0
Gleditsia sinensis	Fabaceae	76	6.61E-10	9
Bupleurum scorzonerifolium	Apiaceae	15	1.07E-08	3
Ardisia japonica	Primulaceae	12	1.68E-08	2
Achyranthes bidentata	Amaranthaceae	8	4.17E-08	4
Sanguisorba officinalis	Rosaceae	27	4.99E-07	14
Cimicifuga foetida	Ranunculaceae	36	7.99E-07	7
Arnebia guttata	Boraginaceae	14	7.99E-07	0
Diphylleia sinensis Li	Berberidaceae	16	1.88E-06	0
Erysimum cheiranthoides L.	Brassicaceae	11	1.88E-06	0
Cimicifuga dahurica	Ranunculaceae	14	6.07E-06	0
Asparagus curillus	Asparagaceae	9	7.63E-06	0
Rubus chingii	Rosaceae	8	7.63E-06	0
Podophyllum emodll	Berberidaceae	15	1.15E-05	0
Lithospermum erythrorhizon	Boraginaceae	18	1.70E-05	2
Strophanthus divaricatus	Apocynaceae	22	2.32E-05	0
Panax ginseng C. A. Mey.	Araliaceae	15	3.09E-05	31
Aralia elata (Miq.) Seem.	Araliaceae	15	1.46E-04	8
Potentilla chinensis	Rosaceae	17	1.51E-04	1
Nerium indicum Mill.	Apocynaceae	9	2.04E-04	1
Paris polyphylla Smith	Melanthiaceae	16	3.33E-04	47
Annona reticulata L.	Annonaceae	66	3.33E-04	1
Phytolacca Americana	Phytolaccaceae	28	5.42E-04	2
Boehmeria nivea	Urticaceae	36	7.12E-04	1
Conyza blinii Levi.	Asteraceae	31	7.75E-04	0
Cestrum nocturnum	Solanaceae	104	2.53E-03	0
Brucea javanica	Simaroubaceae	84	2.55E-03	22
Kochia scoparia	Amaranthaceae	53	3.34E-03	2
Baileya multiradiata	Asteraceae	14	4.70E-03	0
Arnebia euchroma	Boraginaceae	9	4.70E-03	2
Thalictrum minus L.	Ranunculaceae	13	6.24E-03	0
Inula japonica Thunb.	Asteraceae	18	8.21E-03	2
Akebia quinata	Lardizabalaceae	19	8.21E-03	10
Onosma paniculatum	Boraginaceae	41	8.21E-03	2
Lilium brownii	Liliaceae	37	8.30E-03	0
Eupatorium semiserratum	Asteraceae	25	8.30E-03	0
Corydalis incisa	Papaveraceae	14	8.30E-03	0
Eriobotrya japonica	Rosaceae	14	1.22E-02	3
Oplopanax elatus Nakai	Araliaceae	36	1.55E-02	1
Bupleurum falcatum	Apiaceae	17	2.62E-02	1
Aralia taibaiensis	Araliaceae	16	2.62E-02	0
Amaryllis belladonna	Amaryllidaceae	40	3.36E-02	0
Eupatorium rotundifolium	Asteraceae	66	3.36E-02	0
Bupleurum smithii Wolff	Apiaceae	15	3.36E-02	0

Table 1. The predicted anti-cancer plants. The third column represents the number of compounds with anticancer activity in this plant. The last column represents the number of literature and patent whose titles contain both words "the name of plant" and "cancer".

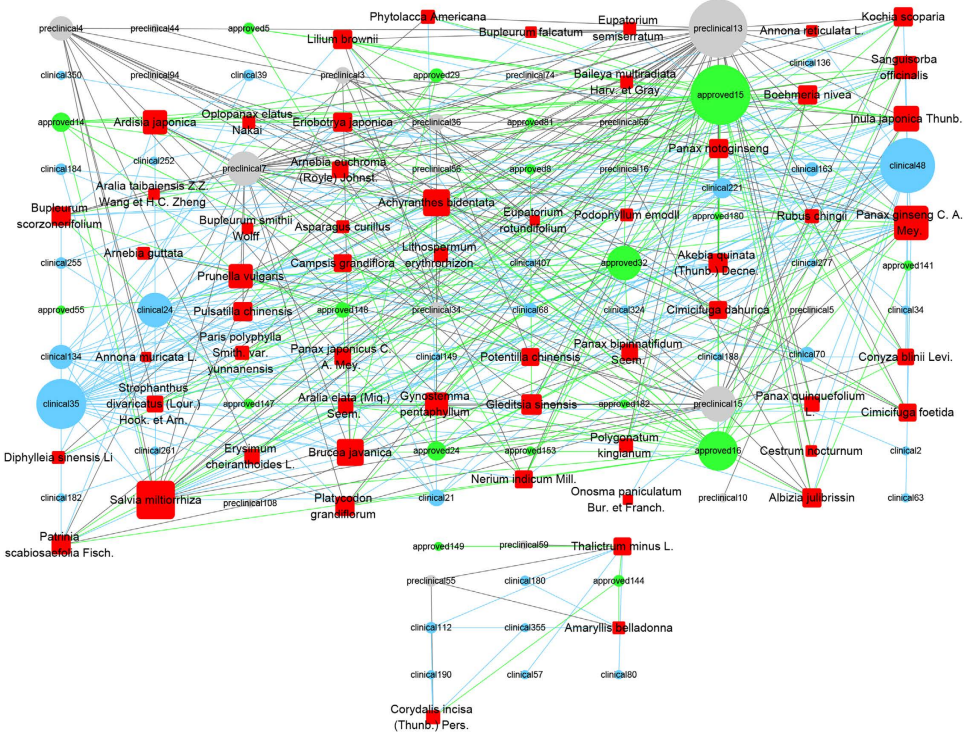

Figure 2. Network of predicted anti-cancer plants and anti-cancer drugs . The network connects plant and drug if the compounds in this plant show similarity with this drug (Tc 0.70, MACCS fingerprint). Two isolated sub-networks were shown in the figure. The red rectangle, green circle, light blue circle and gray circle represent predicted anti-cancer plant, approved drug, clinical drug and preclinical drug, respectively. The lines link the approved drug, clinical drug and preclinical drug are color as green, light blue and gray, respectively. The node size is proportional to the number of connections.

anti-cancer drugs in preclinical, clinical and approved stages from the database of Thomson Reuters Integrity. With the frequent use of anti-cancer drugs and increased duration of treatment, cancer cell may be resistant to the drugs. The problem of drug resistance can be shoveled by developing new and effective anti-cancer drugs. Therefore, these structurally dissimilar compounds are promising molecules and can be used to develop innovative drugs.

Lipinski's rule is often used to determine if a chemical compound with a certain pharmacological activity has properties that would make it a likely orally active drug in humans. The rule evaluates drug-likeness by using four molecular properties (ALogP, molecular weight, H-bond acceptors, and H-bonds donors). The analysis of molecular properties revealed that the distributions of ALogP, molecular weight, H-bond acceptors, and H-bonds donors are very similar and overlapped between the predicted active compounds and cancer drugs. The distribution of rotatable bonds is also similar between the two classes of compounds. These results suggested that most of the predicted active compounds have a good drug-likeness. However, we found that the frequency of most common fragments is very different between the two classes of compounds. Both fragment analysis and molecular property analysis revealed that the ratio of rings and aromatic rings tended to become smaller from the predicted active compounds to cancer drugs. Saturated rings are enriched in the predicted active compounds and unsaturated rings are enriched in the cancer drugs. Generally, unsaturated compounds are more reactive than saturated compounds[45]. Therefore, the reactivity of the predicted active compounds may be lower compared with the cancer drugs. As the degree of reactivity links the level of toxic side effect[46], our results suggested the lower toxicity of the predicted active compounds. In addition, trifluoroethane fragment, a toxic substance, is common in the cancer drugs but absent in the predicted active compounds. It also suggested the lower toxicity of the predicted active compounds.

In our study, we identified 57 anti-cancer plants using the ACEA method which based on the enrichment of anti-cancer compounds in corresponding plant. Literature survey showed that many of these plants have been reported to have anti-cancer activity in several studies, such as Salvia miltiorrhiza, Paris polyphylla, Gynostemma pentaphyllum, Panax ginseng, Panax notoginseng, Brucea javanica, Platycodon grandiflorum. Notably, there are 24 identified anti-cancer plants which were little studied before. Of these plants, 14 plants belong to the families in which many species have already been reported as anti-cancer plants. In contrast, the other 10 plants belong

to the families in which only a few species have been studied as anti-cancer plants, such as caprifoliaceae, solanaceae, bignoniaceae, brassicaceae. The identified plants are widely distributed in 46 genera and 28 families. The identification of these genera and families provides a broader scope and vision for the screening of anti-cancer drugs. These new identified anti-cancer plants are worthy of further studies and provide more chances for the development of cancer drug. Our results may contribute to decision-making in the process of introduction, protection and utilization of medicinal plants. This information of the anti-cancer plants can improve the rationality of decision-making about introduction of medicinal plants.

The prediction of anti-cancer plants requires the annotation information of plant and the compounds in corresponding plant. Incomplete information may affect the results of prediction. For example, there are close to half of 2402 plans which have less than 5 compounds annotated in corresponding plant. Therefore, these plants can not be identified using the ACEA method. Our study mainly based on the TCM Database@Taiwan, which is currently the world's largest and most comprehensive TCM database. With the increasing information in database, the predicted results will be more accurate.

After generation of the plants-drugs network, we found two isolated sub-networks in the overall network. The two sub-networks may be involved in different molecular mechanism of anti-cancer due to connecting different drugs. The smaller sub-network contains two approved drugs (approved144: homoharringtonine and approved149: bosutinib). The bigger sub-network contains 16 approved drugs. In order to probe the molecular mechanisms, we got the target information of these drugs from DrugBank. We found the drugs in the smaller network can bind to the ribosome and inhibit polypeptide chain elongation, thus inhibit protein synthesis. In contrast, the drugs in the bigger network are mainly involved in two molecular mechanism. One is regulation of nuclear receptors and estrogen-related signal. The other is inhibition of DNA replication. Therefore, this result suggests that medicinal plants may exert anti-cancer activity by different molecular mechanism. The plants-drugs network can be used for exploration of molecular mechanism of anti-cancer.

With the accumulation of biological data and increase of the variety and complexity of data types, bioinformatics and cheminformatics play an important role in the integration of these data. Until now, there are two types of data are useful and available for data-mining biologically active compound. One is experimental biological activity data including high-throughput chemical biology screening datasets in Pubchem database[47], such as anti-cancer biological activity data, anti-HIV biological activity data and anti-tuberculosis biological activity data. The other is the curated data about TCM plants and their derived ingredients in several TCM database. The two types of data offer a new opportunity to mine for potential compounds with various activities by using bioinformatics and cheminformatics[48–50]. Salma *et al.* identified anti-tubercular compounds from TCM by integrating anti-tuberculosis biological activity data and TCM related data[50]. Kenneth *et al.* identified quinone subtypes effective against melanoma and leukemia cell by data-mining the GI50 values of the NCI cancer cell line compound[51]. Thomas *et al.* used random forest to virtual screen Chinese herbs for potential inhibitors against several therapeutically important molecular targets[52].

In summary, our analysis suggests that the predicted compounds and plants from TCM database offer an attractive starting point and a broader scope to mine for potential anti-cancer agents. We hope that this study would accelerate in-depth analysis and discovery of anti-cancer agents from TCM.

Methods

To infer anti-cancer plants, we first collected the information concerning the plants and the plant-derived compounds from the TCM Database@Taiwan. The relationship of the pant and its derived compounds was also collected. All compounds were downloaded as mol2 (3D) format. The format was converted to SMILES string[53] by the Open Babel toolbox[54]. A total of 2402 plants and 21334 compounds were collected and downloaded for further study. Detailed information concerning the plants and all compounds can be found in Supplementary **Dataset1** Table S1.

The anti-cancer activities of all the compounds were predicted using CDRUG, which was developed by our laboratory[31]. CDRUG uses a novel molecular description method (relative frequency-weighted fingerprint) and a hybrid score to measure the similarity between the query and the active compounds. Then a confidence level (P-value) is calculated to predict whether a compound has anti-cancer activity. The performance analysis shows that CDRUG has the area under curve of 0.878 and can hit 65% positive results at the false-positive rate of 0.05. Thus CDRUG is effective to predict anti-cancer activity of the chemical compounds. In this study, we used the default ($P < 0.05$) cutoff in CDRUG to screen the 21334 compounds in the TCM Database@Taiwan.

After anti-cancer activity prediction of the 21334 compounds, we measured whether a plant has potential ability to kill cancer cells using the method named ACEA[32]. ACEA is based on the results of anti-cancer activity prediction and uses a hypergeometric distribution to perform enrichment analysis. The P-value of each plant can be calculated using the following equation:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \tag{1}$$

Here, N and n are the total number of compounds and the total number of anti-cancer compounds in the TCM Database@Taiwan, respectively; m and k represent the number of compounds and the number of anti-cancer compounds in a plant, respectively. Both n and k are calculated using CDRUG. Because multiple tests (2402 plants) were performed, the Bonferroni correction method was used to adjust the P-value determined by ACEA:

$$P\_adj = p \times Ng \tag{2}$$

Here, P_adj is the adjusted P-value of ACEA, P is the P-value of ACEA (without Bonferroni correction) and Ng is the number of plants in the TCM Database@Taiwan. Only plants with P_adj < 0.05 were retained.

In order to compare the similarity of the predicted anti-cancer compounds with the anti-cancer drugs in the different development stages, we got the information concerning the anti-cancer drugs in preclinical, clinical and approved stages from the database of Thomson Reuters Integrity (www.thomsonreutersintegrity.com). The molecular properties of the predicted active compounds and anti-cancer drugs were calculated using the protocol 'Calculate Molecular Properties' in Pipeline Pilot v8.5[55]. The calculated properties include ALogP, molecular weight, and the number of rotatable bonds, rings, aromatic rings, H-bond acceptors, and H-bonds donors, and so on. Detailed information and molecular properties for the predicted active compounds and anti-cancer drugs can be found in Supplementary **Dataset1** Table S2. The most common fragments and their frequency were calculated using the protocol 'Most Frequent Fragments' Pipeline Pilot v8.5. These fragments and their frequency are available in Supplementary **Dataset1** Table S4. The structural similarity was measured by Tanimoto coefficient (Tc)[56]. Tc is defined as $Tc = C(i, j)/U(i, j)$, where C(i, j) is the number of common features in the fingerprints of molecules i and j and where U(i, j) is the number of all features in the union of the fingerprints of molecules i and j. The fingerprint MACCS implemented in the Pybel[57] were generated for each structure and used to calculate TC. Two compounds are considered structurally similar if their fingerprints have a Tc of 0.70 or greater[58,59]. After calculation, the similarity network was visualized using Cytoscape v3.2[60].

References

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *cell* **144,** 646–674 (2011).
2. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *cell* **100,** 57–70 (2000).
3. Organization, W. H. *Cancer: Fact sheet N297*, February 2015. *URL* http://www.who.int/mediacentre/factsheets/fs297/en/ (2015).
4. Cancer, I. A. f. R. o. World cancer report 2014. *Geneva: WHO* (2014).
5. Thun, M. J., DeLancey, J. O., Center, M. M., Jemal, A. & Ward, E. M. The global burden of cancer: priorities for prevention. *Carcinogenesis* **31,** 100–110, doi: 10.1093/carcin/bgp263 (2010).
6. Safarzadeh, E., Sandoghchian Shotorbani, S. & Baradaran, B. Herbal medicine as inducers of apoptosis in cancer treatment. *Advanced pharmaceutical bulletin* **4,** 421–427, doi: 10.5681/apb.2014.062 (2014).
7. Qi, F. *et al.* Chinese herbal medicines as adjuvant treatment during chemo-or radio-therapy for cancer. *Biosci Trends* **4,** 297–307 (2010).
8. Pereira, D. M., Valentao, P., Correia-da-Silva, G., Teixeira, N. & Andrade, P. B. Plant Secondary Metabolites in Cancer Chemotherapy: Where are We? *Current Pharmaceutical Biotechnology* **13,** 632–650 (2012).
9. Coseri, S. Natural products and their analogues as efficient anticancer drugs. *Mini reviews in medicinal chemistry* **9,** 560–571 (2009).
10. Tascilar, M., de Jong, F. A., Verweij, J. & Mathijssen, R. H. Complementary and alternative medicine during cancer treatment: beyond innocence. *The oncologist* **11,** 732–741 (2006).
11. Wang, C.-Z., Calway, T. & Yuan, C.-S. Herbal medicines as adjuvants for cancer therapeutics. *The American journal of Chinese medicine* **40,** 657–669 (2012).
12. Cragg, G. M. & Newman, D. J. Plants as a source of anti-cancer agents. *Journal of Ethnopharmacology* **100,** 72–79, doi: 10.1016/j.jep.2005.05.011 (2005).
13. Graham, J. G., Quinn, M. L., Fabricant, D. S. & Farnsworth, N. R. Plants used against cancer - an extension of the work of Jonathan Hartwell. *Journal of Ethnopharmacology* **73,** 347–377, doi: 10.1016/s0378-8741(00)00341-x (2000).
14. Chen, C. Y. TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. *PloS one* **6,** e15939, doi: 10.1371/journal.pone.0015939 (2011).
15. Wang, S., Penchala, S., Prabhu, S., Wang, J. & Huang, Y. Molecular basis of traditional Chinese medicine in cancer chemoprevention. *Current drug discovery technologies* **7,** 67–75 (2010).
16. Han, J. Traditional Chinese medicine and the search for new antineoplastic drugs. *J Ethnopharmacol* **24,** 1–17 (1988).
17. Yang, G. *et al.* Traditional chinese medicine in cancer care: a review of case series published in the chinese literature. *Evidence-based complementary and alternative medicine: eCAM* **2012,** 751046, doi: 10.1155/2012/751046 (2012).
18. Konkimalla, V. B. & Efferth, T. Evidence-based Chinese medicine for cancer therapy. *J Ethnopharmacol* **116,** 207–210, doi: 10.1016/j.jep.2007.12.009 (2008).
19. Normile, D. Asian medicine. The new face of traditional Chinese medicine. *Science* **299,** 188–190, doi: 10.1126/science.299.5604.188 (2003).
20. Ouyang, L. *et al.* Plant natural products: from traditional compounds to new emerging drugs in cancer therapy. *Cell Proliferation* **47,** 506–515, doi: 10.1111/cpr.12143 (2014).
21. Liu, J., Ouyang, L., Chen, Y. & Liu, B. Plant natural compounds targeted cancer cell autophagy: research advances. *Journal of International Pharmaceutical Research* **40,** 688–694 (2013).
22. Wang, H. *et al.* Plants vs. Cancer: A Review on Natural Phytochemicals in Preventing and Treating Cancers and Their Druggability. *Anti-Cancer Agents Med. Chem.* **12,** 1281–1305 (2012).
23. Grohs, B. M. *et al.* Plant-Produced Trastuzumab Inhibits the Growth of HER2 Positive Cancer Cells. *Journal of Agricultural and Food Chemistry* **58,** 10056–10063, doi: 10.1021/jf102284f (2010).
24. Efferth, T. Cancer Therapy with Natural Products and Medicinal Plants. *Planta Medica* **76,** 1035–1036, doi: 10.1055/s-0030-1250062 (2010).
25. Suh, Y., Afaq, F., Johnson, J. J. & Mukhtar, H. A plant flavonoid fisetin induces apoptosis in colon cancer cells by inhibition of COX2 and Wnt/EGFR/NF-kappa B-signaling pathways. *Carcinogenesis* **30,** 300–307, doi: 10.1093/carcin/bgn269 (2009).
26. Loa, J., Chow, P. & Zhang, K. Studies of structure-activity relationship on plant polyphenol-induced suppression of human liver cancer cells. *Cancer Chemotherapy and Pharmacology* **63,** 1007–1016, doi: 10.1007/s00280-008-0802-y (2009).
27. Newman, D. J. Natural products as leads to potential drugs: An old process or the new hope for drug discovery? *J Med Chem* **51,** 2589–2599, doi: 10.1021/jm0704090 (2008).
28. Kaur, P., Shukla, S. & Gupta, S. Plant flavonoid apigenin inactivates Akt to trigger apoptosis in human prostate cancer: an *in vitro* and *in vivo* study. *Carcinogenesis* **29,** 2210–2217, doi: 10.1093/carcin/bgn201 (2008).
29. Efferth, T., Li, P. C., Konkimalla, V. S. & Kaina, B. From traditional Chinese medicine to rational cancer therapy. *Trends in molecular medicine* **13,** 353–361, doi: 10.1016/j.molmed.2007.07.001 (2007).
30. Konkimalla, V. B. & Efferth, T. Anti-cancer natural product library from traditional chinese medicine. *Combinatorial chemistry & high throughput screening* **11,** 7–15 (2008).
31. Li, G. H. & Huang, J. F. CDRUG: a web server for predicting anticancer activity of chemical compounds. *Bioinformatics* **28,** 3334–3335, doi: 10.1093/bioinformatics/bts625 (2012).
32. Li, G. H. & Huang, J. F. Inferring therapeutic targets from heterogeneous data: HKDC1 is a novel potential therapeutic target for cancer. *Bioinformatics* **30,** 748–752, doi: 10.1093/bioinformatics/btt606 (2014).

33. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nature reviews. Cancer* **6,** 813–823, doi: 10.1038/nrc1951 (2006).
34. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews* **64,** 4–17 (2012).
35. Wang, N. *et al.* A polysaccharide from Salvia miltiorrhiza Bunge improves immune function in gastric cancer rats. *Carbohydrate Polymers* **111,** 47–55, doi: 10.1016/j.carbpol.2014.04.061 (2014).
36. Hu, T. *et al.* Reversal of P-glycoprotein (P-gp) mediated multidrug resistance in colon cancer cells by cryptotanshinone and dihydrotanshinone of Salvia miltiorrhiza. *Phytomedicine* **21,** 1264–1272, doi: 10.1016/j.phymed.2014.06.013 (2014).
37. Lee, W. Y. W. *et al.* Cytotoxic Effects of Tanshinones from Salvia miltiorrhiza on Doxorubicin-Resistant Human Liver Cancer Cells. *Journal of Natural Products* **73,** 854–859, doi: 10.1021/np900792p (2010).
38. Gong, Y. *et al.* Bioactive tanshinones in Salvia Miltiorrhiza inhibit the growth of prostate cancer cells *in vitro* and in mice. *International Journal of Cancer* **129,** 1042–1052, doi: 10.1002/ijc.25678 (2011).
39. Wang, X. H. *et al.* Antitumor agents. 239. isolation, structure elucidation, total synthesis, and anti-breast cancer activity of neotanshinlactone from Salvia miltiorrhiza. *J Med Chem* **47,** 5816–5819, doi: 10.1021/jm040112r (2004).
40. Li, F.-R. *et al.* Paris polyphylla Smith Extract Induces Apoptosis and Activates Cancer Suppressor Gene Connexin26 Expression. *Asian Pacific Journal of Cancer Prevention* **13,** 205–209, doi: 10.7314/apjcp.2012.13.1.205 (2012).
41. Lee, M. S. *et al.* Effects of polyphyllin D, a steroidal saponin in Paris polyphylla, in growth inhibition of human breast cancer cells and in xenograft. *Cancer Biology & Therapy* **4,** 1248–1254, doi: 10.4161/cbt.4.11.2136 (2005).
42. Huang, Y. *et al.* Separation and identification of steroidal compounds with cytotoxic activity against human gastric cancer cell lines *in vitro* from the rhizomes of Paris polyphylla var. chinensis. *Chemistry of Natural Compounds* **43,** 672–677, doi: 10.1007/s10600-007-0225-8 (2007).
43. He, H., Sun, Y.-P., Zheng, L. & Yue, Z.-G. Steroidal saponins from Paris polyphylla induce apoptotic cell death and autophagy in A549 human lung cancer cells. *Asian Pacific journal of cancer prevention: APJCP* **16,** 1169–1173 (2015).
44. Sultana, S. *et al.* Medicinal Plants Combating Against Cancer - a Green Anticancer Approach. *Asian Pacific Journal of Cancer Prevention* **15,** 4385–4394, doi: 10.7314/apjcp.2014.15.11.4385 (2014).
45. Bergman, R. G. Organometallic chemistry: C–H activation. *Nature* **446,** 391–393 (2007).
46. Barrett, D. Proteinase and Peptidase Inhibition: Recent Potential Targets for Drug Development. *Drug Discovery Today* **7,** 1124 (2002).
47. Wang, Y. *et al.* PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **37,** W623–W633 (2009).
48. Li, X.-J., Kong, D.-X. & Zhang, H.-Y. Chemoinformatics approaches for traditional Chinese medicine research and case application in anticancer drug discovery. *Curr. Drug Discovery Technol.* **7,** 22–31 (2010).
49. Zhang, K., Li, Y., Zhang, Z., Guan, W. & Pu, Y. [Chemoinformatics study on antibacterial activity of traditional Chinese medicine compounds]. *China J. Chin. Mater. Med.* **38,** 777–780 (2013).
50. Jamal, S. & Scaria, V. Data-mining of potential antitubercular activities from molecular ingredients of traditional Chinese medicines. *PeerJ* **2,** e476 (2014).
51. Marx, K. A., O'Neil, P., Hoffman, P. & Ujwal, M. Data mining the NCI cancer cell line compound GI50 values: identifying quinone subtypes effective against melanoma and leukemia cell classes. *J. Chem Inf. Model.* **43,** 1652–1667 (2003).
52. Ehrman, T. M., Barlow, D. J. & Hylands, P. J. Virtual screening of Chinese herbs with random forest. *J. Chem Inf. Model.* **47,** 264–278 (2007).
53. Weininger, D. S. M. I. L. E. S., a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **28,** 31–36 (1988).
54. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *Journal of cheminformatics* **3,** 33, doi: 10.1186/1758-2946-3-33 (2011).
55. Pilot, P. Version 8.5. Accelrys. *Inc. San Diego, CA* **92121** (2011).
56. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **11,** 1046–1053, doi: 10.1016/j.drudis.2006.10.005 (2006).
57. O'Boyle, N. M., Morley, C. & Hutchison, G. R. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chemistry Central journal* **2,** 5, doi: 10.1186/1752-153X-2-5 (2008).
58. Peltason, L. & Bajorath, J. Molecular similarity analysis uncovers heterogeneous structure-activity relationships and variable activity landscapes. *Chemistry & biology* **14,** 489–497, doi: 10.1016/j.chembiol.2007.03.011 (2007).
59. Zhong, S. *et al.* Identification and validation of human DNA ligase inhibitors using computer-aided drug design. *J Med Chem* **51,** 4553–4562, doi: 10.1021/jm8001668 (2008).
60. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13,** 2498–2504, doi: 10.1101/gr.1239303 (2003).

Acknowledgements

This work was supported by the National Basic Research Program of China (Grant No. 2013CB835100), the Instruments Function Deployment Foundation of CAS (Grants Nos yg2010044, yg2011057 and 2014gk01), and the National Natural Science Foundation of China (Grant No. 31401142 to D.S.X., No. 31401137 to G.H.L. and No. 31123005 to J.F.H.).

Author Contributions

S.-X.D., G.-H.L. and J.-F.H. participated in research design. S.-X.D., W.-X.L., F.-F.H., Y.-C.G., J.-J.Z., J.-Q.L., Q.W., Y.-D.G. performed data analysis. S.-X.D., G.-H.L. and J.-F.H. wrote or contributed to the writing of the manuscript. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Dai, S.-X. *et al.* In silico identification of anti-cancer compounds and plants from traditional Chinese medicine database. *Sci. Rep.* **6**, 25462; doi: 10.1038/srep25462 (2016).