D1110–D1117 Nucleic Acids Research, 2019, Vol. 47, Database issue Published online 31 October 2018
doi: 10.1093/nar/gky1021

SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping

Yang Wu[1,2,†], Feilong Zhang[1,†], Kuo Yang[3,†], Shuangsang Fang[2], Dechao Bu[2], Hui Li[2], Liang Sun[2], Hairuo Hu[2], Kuo Gao[1], Wei Wang[1], Xuezhong Zhou[3,*], Yi Zhao[1,2,*] and Jianxin Chen[1,*]

[1]Beijing University of Chinese Medicine, ChaoYang District, Beijing 100029, China, [2]Key Laboratory of Intelligent Information Processing, Advanced Computer Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China and [3]School of Computer and Information Technology and Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China

Received August 15, 2018; Revised September 25, 2018; Editorial Decision October 11, 2018; Accepted October 22, 2018

ABSTRACT

Recently, the pharmaceutical industry has heavily emphasized phenotypic drug discovery (PDD), which relies primarily on knowledge about phenotype changes associated with diseases. Traditional Chinese medicine (TCM) provides a massive amount of information on natural products and the clinical symptoms they are used to treat, which are the observable disease phenotypes that are crucial for clinical diagnosis and treatment. Curating knowledge of TCM symptoms and their relationships to herbs and diseases will provide both candidate leads and screening directions for evidence-based PDD programs. Therefore, we present SymMap, an integrative database of traditional Chinese medicine enhanced by symptom mapping. We manually curated 1717 TCM symptoms and related them to 499 herbs and 961 symptoms used in modern medicine based on a committee of 17 leading experts practicing TCM. Next, we collected 5235 diseases associated with these symptoms, 19 595 herbal constituents (ingredients) and 4302 target genes, and built a large heterogeneous network containing all of these components. Thus, SymMap integrates TCM with modern medicine in common aspects at both the phenotypic and molecular levels. Furthermore, we inferred all pairwise relationships among SymMap components using statistical tests to give pharmaceutical scientists the ability to rank and filter promising results to guide drug discovery. The SymMap database can be accessed at http://www.symmap.org/ and https://www.bioinfo.org/symmap.

INTRODUCTION

Two main approaches are used in modern drug discovery: target-based drug discovery (TDD) and phenotypic drug discovery (PDD) (1). TDD begins with a well-defined molecular target for a specific disease, and compound libraries are generated from which optimal compounds with activity against the target are identified. In contrast, PDD does not rely on knowledge of molecular targets, but is rather based on screening a large number of compounds and monitoring phenotypic changes. An influential analysis in 2011 reported that PDD has been more productive than TDD as a means of discovering first-in-class drugs (2).

Isolation and further derivatization of natural products from traditional medicines is a promising PDD strategy (3). The traditional use of natural products has been extensively documented in diverse cultures for millennia, and these descriptions provide valuable therapeutics drug leads for specific disease phenotypes. It is shown that, of 122 traditional medicine-derived compounds used as drugs in countries hosting WHO-Traditional Medicine Centers, 80% were used for their traditional purpose or a related ethnomedical purpose (4). These findings demonstrate the value of traditional medicinal knowledge in the quest to discover new biologically active compounds.

Traditional Chinese medicine (TCM) provides a massive amount of information on natural products and the clinical symptoms they are used to treat (5), which are the observable disease phenotypes that are crucial for clinical diagnosis and treatment (6). These empirical knowledge can shed light on PDD screening directions in modern drug discovery. For example, the discovery of ephedrine, an anti-

[*]To whom correspondence should be addressed. Tel: +86 10 6260 0822; Fax: +86 10 6260 1356; Email: biozy@ict.ac.cn
Correspondence may also be addressed to Xuezhong Zhou. Tel: +86 10 5168 4931; Fax: +86 10 5168 4931; Email: xzzhou@bjtu.edu.cn
Correspondence may also be addressed to Jianxin Chen. Tel: +86 10 6428 6398; Fax: +86 10 6428 6398; Email: cjx@bucm.edu.cn
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint first authors.

asthmatic drug identified by the first TCM pharmacologist Kehui Chen (1898–1988), was inspired by the clinical use of the Chinese herb *Ma Huang* to treat asthma for >4000 years (7). Artemisinin (*qinghaosu*), the first-line drug for malaria, was discovered by 2015 Nobel laureate Youyou Tu, who was inspired by the Chinese herb *qinghao* for combating the symptoms of malaria in TCM (8). Consequently, standardization of the symptom vocabulary of TCM and further illustration of the relationships between symptoms, natural products (mainly herbs), diseases, and molecular targets has the potential to provide novel lead/drug candidates.

Knowledge of the symptoms traditionally treated by herbs is difficult for modern pharmaceutical scientists to understand for two reasons. Firstly, most TCM symptom terms are written in ancient Chinese. However, only a tiny fraction of Chinese intellectuals alive today can understand the definitions of TCM symptoms exactly. Secondly, to better leverage the knowledge of TCM usage, TCM symptoms must be mapped onto the terms for symptoms used in modern medicine (MM). As TCM is based on a holistic philosophy that differs substantially from that of MM (9), this task can be accomplished only by experts who are trained in TCM and familiar with MM. However, the number of individuals qualified to perform this task has declined continuously in recent years. Accordingly, linking TCM symptom-herb relationships to MM, as well as the molecular mechanisms underlying diseases, is urgent.

Therefore, we built a new database, SymMap, an integrative database of traditional Chinese medicine enhanced by symptom mapping. During the development of SymMap, the difficulties mentioned above were overcome by forming a committee of 17 leading experts practicing TCM. SymMap provides four types of new knowledge. Firstly, we manually standardized TCM symptom terms and definitions, which were mapped to herbs registered in the Chinese Pharmacopoeia, the collection of TCM knowledge with very high level of evidence. Secondly, we rigorously mapped these TCM symptoms to MM symptoms recorded in the unified medical language system (UMLS) via expert consensus and subsequent manual verification. Thirdly, using database mining, we mapped the knowledge of symptom-herb relationships onto current data regarding the molecular mechanisms of TCM, including the compound compositions of herbs (ingredients), their molecular targets (mainly genes/proteins), and diseases related to symptoms or targets. Finally, we present all-versus-all pairwise associations among all components in SymMap, with some of them analyzed by statistical inference to enable pharmaceutical scientists to rank and filter the most promising results.

In the last decade, several databases focusing on different aspects of TCM knowledge have been published. For example, the TCM-ID (10), HIT (11), TCMID (12), and TCMSP (13) databases. These databases have undergone continuous stepwise improvement as new components or aspects of TCM have been added. However, information regarding symptoms and phenotypes has never been curated, standardized, and connected to herbs and diseases, as well their underlying molecular mechanisms. Thus, SymMap fills the gap and presents the newly curated symptom-herb knowledge, which can provide both pharmacological effects (phenotypic changes) and candidate leads for PDD screening efforts. In addition, SymMap provides symptom-mechanism mapping that will enable further analysis of the shared symptoms and targets of multiple diseases for accelerating drug repositioning studies.

DATA COLLECTION AND PROCESSING

Data sources of SymMap

SymMap contains six components: symptoms used in TCM (TCM symptoms) and MM (MM symptoms), herbs, ingredients, targets (also denoted as genes in the article) and diseases (Figure 1A). Among these components, TCM symptoms, MM symptoms and herbs can be regarded as the phenotypic knowledge that are valuable for PDD programs, whereas the ingredients, targets, and diseases consist of molecular information derived from TDD efforts.

We introduced phenotype-level information by extracting TCM symptoms and herb terms from the Chinese Pharmacopoeia (CHPH, 2015 edition). A description about the fields of each record in CHPH is illustrated in Supplementary Figure S1. We firstly invited 17 leading experts practicing TCM (Supplementary Table S1) to manually check all symptom terms from the CHPH (Figure 1B). The names of TCM symptoms were standardized according to an authoritative TCM publication, 'Standardization research on TCM Terminology' (*Zhongyiyao Mingci Shuyu Guifanhua Yanjiu*, published in 2016), and a published platform for integrating TCM terminologies (14). And we curated the definition, locus, and property information for these symptoms using another TCM publication, 'Standardization of Pathological Terminology' (*Bingzhuang Shuyu Guifanhua Jichu*, published in 2015). Then, we collected MM symptom terms from the MeSH (version 2017) (15), SIDER (version 2017) (16) and UMLS (version 2016) (17) databases, after which the expert committee manually mapped TCM symptoms to MM symptoms.

Next, we collected molecular information mainly by database integration. For example, we integrated the ingredient information from the TCMID (version 2015), TCMSP (version 2.3) and TCM-ID (version 1.0) databases. To obtain non-redundant ingredient records for herbs, we select from different records with common identifier provided by the three databases, such as CAS, PubChem_CID and InChI Key etc. The target component of the database was collected from two sources: the target genes of TCM compounds from the HIT (version 2.0) and TCMSP databases, and the target genes of modern diseases from the HPO (version 2017) (18), DrugBank (version 5.0.0) (19) and NCBI gene (version 2018) (20) databases. Similarly, the disease component was also merged from two sources, namely OMIM (version 2017) (21) and Orphanet (version 2017) (22) databases. The sources of all components of the SymMap database are summarized in Table 1.

Direct associations among SymMap components

There are six direct associations among the components (Figure 1A). Two relationships, the TCM symptom-herb and TCM symptom-MM symptom associations, had never before been included in a public database, but they were

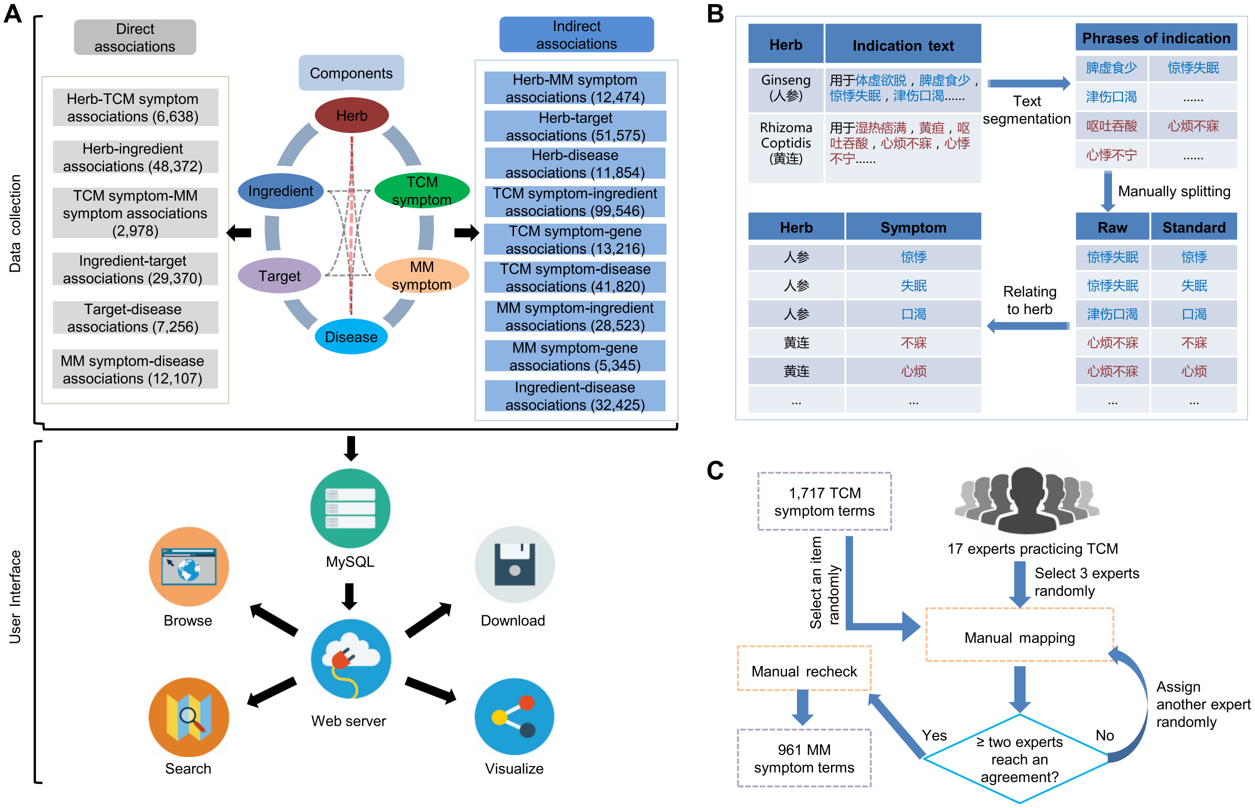

A
Data collection
Direct associations
Herb-TCM symptom associations (6,638)
Herb-ingredient associations (48,372)
TCM symptom-MM symptom associations (2,978)
Ingredient-target associations (29,370)
Target-disease associations (7,256)
MM symptom-disease associations (12,107)
Components
Herb
Ingredient
TCM symptom
Target
MM symptom
Disease
Indirect associations
Herb-MM symptom associations (12,474)
Herb-target associations (51,575)
Herb-disease associations (11,854)
TCM symptom-ingredient associations (99,546)
TCM symptom-gene associations (13,216)
TCM symptom-disease associations (41,820)
MM symptom-ingredient associations (28,523)
MM symptom-gene associations (5,345)
Ingredient-disease associations (32,425)
User Interface
MySQL
Browse
Download
Web server
Search
Visualize
B
Herb
Indication text
Ginseng (人参)
用于体虚欲脱，脾虚食少，惊悸失眠，津伤口渴……
Rhizoma Coptidis (黄连)
用于湿热痞满，黄疸，呕吐吞酸，心烦不寐，心悸不宁……
Text segmentation
Phrases of indication
脾虚食少
惊悸失眠
津伤口渴
……
呕吐吞酸
心烦不寐
心悸不宁
……
Manually splitting
Raw
Standard
惊悸失眠
惊悸
惊悸失眠
失眠
津伤口渴
口渴
心烦不寐
不寐
心烦不寐
心烦
...
...
Relating to herb
Herb
Symptom
人参
惊悸
人参
失眠
人参
口渴
黄连
不寐
黄连
心烦
...
...
C
1,717 TCM symptom terms
Select an item randomly
17 experts practicing TCM
Select 3 experts randomly
Manual mapping
Assign another expert randomly
Manual recheck
961 MM symptom terms
Yes
≥ two experts reach an agreement?
No

Figure 1. Schematic of the SymMap database. (**A**) Upper panel: the six components contained in SymMap are illustrated in six circles in different colors in the middle. The blue arcs connecting the circles show the six direct associations, with the numbers of associations shown at the left. The gray dotted lines connecting the six components show the nine indirect associations, with the numbers of associations shown in the right. Lower panel: implementation of the functions of SymMap. (**B**) Illustration of the scheme for extraction, curation and standardization of TCM symptom terms and their relationships to herbs. (**C**) Illustration of the scheme for expert curation of TCM symptom-MM symptom mapping.

Table 1. Overview of the data curated in SymMap

Components	Data source	Amount
Herbs	Extracted from the Chinese pharmacopoeia (2015 edition)	499
TCM symptoms	Extracted, manually curated, and standardized from the Chinese pharmacopoeia (2015 edition)	1717
MM symptoms	Indexed in the UMLS database, and manually mapped to TCM symptoms	961
Ingredients	Integrated from the TCMID, TCMSP and TCM-ID databases	19 595
Targets	Integrated from the HIT, TCMSP, HPO, DrugBank and NCBI databases	4302
Diseases	Integrated from the OMIM, MeSH and Orphanet databases	5235

included in SymMap for the first time as a result of manual curation by experts. Other relationships, including the MM symptom–disease, herb–ingredient, ingredient–target and target–disease (also referred to as gene–disease) associations, are dispersed distributed in multiple databases that required integration.

We firstly mapped two types of direct associations to TCM symptoms by manual curation. The TCM symptom–herb relationships were obtained directly from the CHPH after standardization of TCM symptom terms (Figure 1B). TCM symptom-MM symptom mapping was conducted using an iterative process. For each TCM symptom term, three experts were randomly selected and given a full list of MM symptom terms. If the experts did not map the TCM term to the same MM symptom then another expert was assigned, and this process was repeated until at least two experts reached an agreement. After all TCM symptoms were mapped, manual rechecking was conducted to ensure the accuracy of the database (Figure 1C). Note that the full list of MM symptoms in UMLS identifiers contains not only concepts about symptoms, but also other types of terms in modern medicine (Supplementary Table S2).

Next, we curated additional direct associations by database integration. The MM symptom–disease relationships were aligned and connected based on the HPO, OMIM and Orphanet databases. We mapped the UMLS ids of MM symptoms into the HPO ids first, and then related the HPO identifiers of symptoms to the disease terms in OMIM or Orphanet identifiers based on the HPO records. It is noteworthy that a number of diseases have both OMIM and Orphanet identifiers. In this case, we merged the disease terms according to their names in a case insensitive way. The herb–ingredient associations were merged from the TCMSP, TCMID, and TCM-ID databases. The ingredient–target associations were obtained from the HIT and TCMSP databases, whereas the gene–disease associations were aligned and obtained from the HPO and OMIM databases. For database integration, we carefully checked the results to make sure that the final lists were non-redundant. The sources of all direct associations are summarized in Supplementary Table S3.

Indirect associations among SymMap components

In addition to six direct associations involving adjacent components, there were nine indirect associations involving non-adjacent components (Figure 1A). We chose to infer indirect associations from combinations of direct relationships. For example, the indirect relationships between herbs and MM symptoms, can be obtained using the TCM symptom as a middle component (Supplementary Figure S2A). To remove possible false positives, we used Fisher's Exact Test (23) to obtain reliable associations with statistical significance, and Fisher's exact test is effective and widely used for evaluating the reliability of biomedical associations (24). Furthermore, to control the false discovery rate (FDR) due to multiple tests, we calculated the FDRs according to both the Bonferroni (25) and the BH (26) methods from P-values. The strategy was used for the inference of four indirect associations that can be connected through a middle component between them. For example, the indirect associations for TCM symptom-ingredient, herb-target, ingredient-disease, and MM symptom-target relationships were inferred through herbs, ingredients, targets, and diseases, respectively (Supplementary Figure S2B-E). Note that for herb-MM symptom and TCM symptom–disease relationships, we did not perform tests, but retained all associations using the intermediates TCM symptom and MM symptom, respectively (Supplementary Figure S2A, F). Because the intermediate relationships manually curated by experts were sufficiently convincing to retain.

For the three remaining indirect associations, which could have been connected by at least two components, it was required that one component was selected as the intermediate. For example, we selected the disease component for the TCM symptom–target indirect associations and applied the same test procedure used above with only one middle component (Supplementary Figure S3A). The only difference between this procedure and that mentioned above was that the statistical inferences of TCM symptom–target relationships were based on the TCM symptom–disease relationship inferred previously, so the strategy had two steps. Similarly, the indirect MM symptom–ingredient associations were inferred in a step-wise manner using TCM symptoms as the intermediate (Supplementary Figure S3B). It is noteworthy that the herb-disease association was emphasized and obtained using three strategies because this relationship is an important guide for PDD (Supplementary Figure S3C). Firstly, we manually curated a small number of herb–disease relationships from the CHPH, as the indications for herbs in CHPH contain a fraction of disease information (Supplementary Figure S1). Secondly, we used ingredients as the intermediate to conduct two-step testing. Thirdly, we used MM symptoms as the intermediate to conduct two-step testing. For herb–disease pairs that were inferred by multiple methods, the smallest *P*-values and FDRs were selected as confidence scores. We should note that the one component chosen as the intermediate is selected based on empirical knowledge. The sources of all indirect associations are summarized in Supplementary Table S4.

Implementation of SymMap

In summary, SymMap provides information about six components related to TCM and MM and their pairwise relationships using a convenient web interface from which users can browse, search, visualize and download data. SymMap is free to access at http://www.symmap.org and https://www.bioinfo.org/symmap without user registration. The SymMap website was built using the Python-Flask and Nginx frameworks. The SymMap data are stored in a MySQL database. The SymMap website is compatible with most major browsers.

DATABASE CONTENTS AND ACCESS

Database statistics

The six curated components in SymMap include 1717 TCM symptoms, 961 MM symptoms, 499 herbs, 19 595 ingredients, 4302 targets and 5235 diseases (Table 1). The six types of direct associations in SymMap include 6638 herb–TCM symptom associations, 2978 TCM symptom–MM symptom associations, 48 372 herb–ingredient associations, 12 107 MM symptom–disease associations, 29 370 ingredient–target associations and 7256 gene–disease associations (Figure 1A). The distributions of the connections for each type of direct association are shown in Figure 2A. For example, in the TCM symptom–herb associations, each herb is associated with 13.30 TCM symptoms on average, and each TCM symptom is associated with 3.87 herbs on average. For the TCM symptom-MM symptom mapping introduced by SymMap, each TCM symptom is associated with 1.74 MM symptoms, and each MM symptom is associated with 3.13 TCM symptoms. The details of all direct associations are shown in Supplementary Table S3.

The compositions of all nine indirect associations are summarized in Supplementary Table S4. As expected, the Bonferroni method for multiple testing correction was quite strict and gave rather small set of predictions. So we mainly chose the result from the BH method for representation.

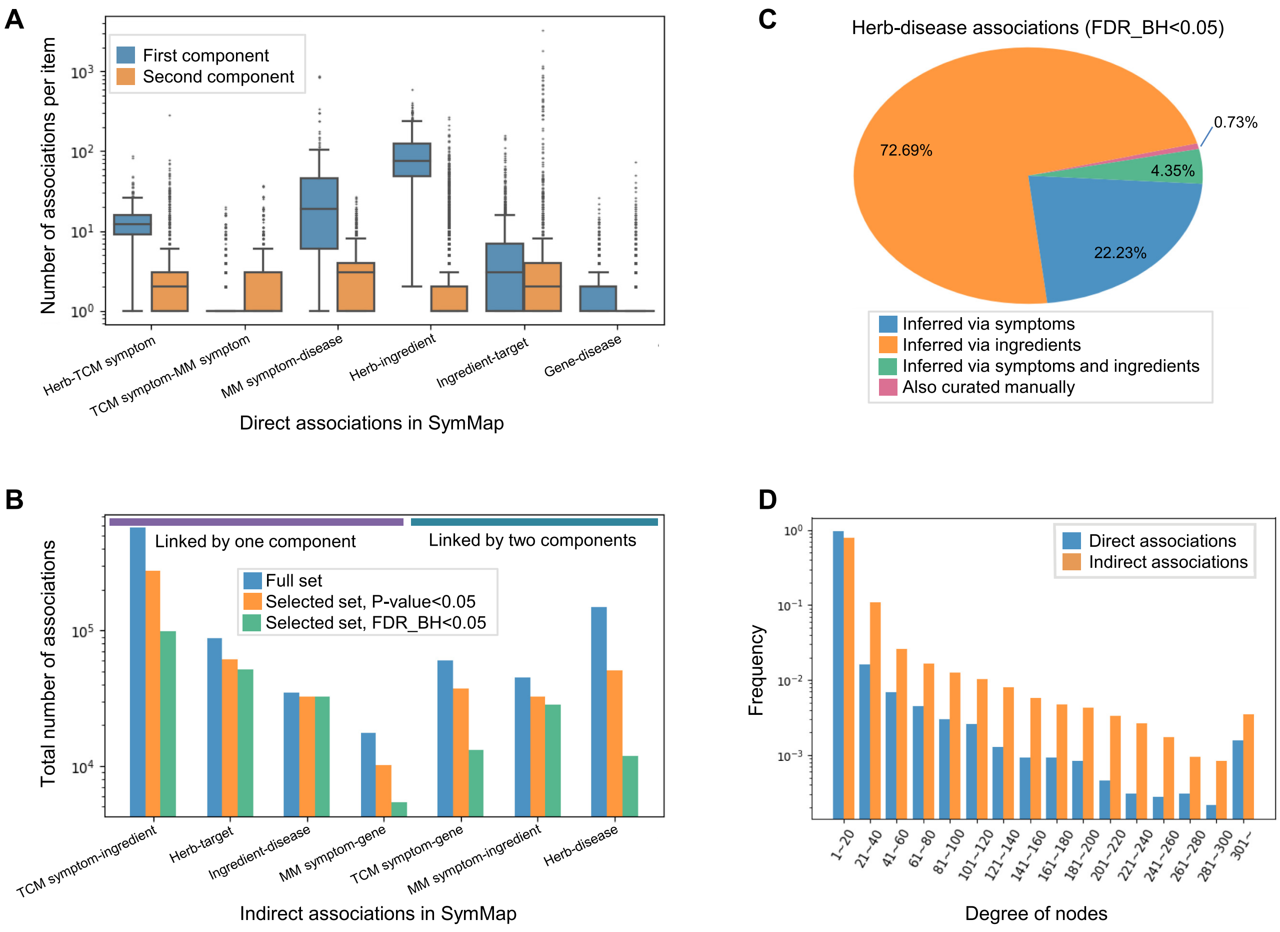


Figure 2. Characteristics of the SymMap integrative network. (**A**) Box plots show the distribution of association numbers per item for six direct associations. For each association between component 1–component 2, two boxes are shown. The first box in blue shows the distribution of component 1, whereas the second box in orange shows the distribution of component 2. (**B**) Bar plots show the total number of associations for seven indirect associations. For each association, three bars are shown. The first bar in blue shows the full set, the second bar in orange show the loosely selected set (P-value <0.05), and the third bar in green shows the stringently selected set (FDR_BH < 0.05). (**C**) The sources of indirect herb-disease associations are shown. Associations inferred via symptoms are shown in blue, whereas those inferred via ingredients are shown in orange, those inferred via both symptoms and ingredients are shown in green, and those also be curated manually are shown in red. (**D**) The distribution of node degrees in the heterogeneous network of SymMap, with direct associations shown in blue and indirect associations shown in orange.

For herb-MM symptom and TCM symptom–disease relationships, we provided all possible associations by network neighbor extension (full set) because no statistical tests were conducted. For the other seven types of indirect association, we compared three datasets, including all possible associations (full set), statistically significant associations with loose criteria (selected set, $P < 0.05$), and statistically significant associations with stringent criteria (selected set, FDR_BH < 0.05) (Figure 2B). The total number of associations was reduced as stricter criteria were adopted. For example, the full set of TCM symptom–ingredient associations contains 576 129 associations, but applying a P-value cut-off of 0.05 leaves 275 097 associations, whereas applying a FDR (BH) cutoff of 0.05 leaves 99 546 associations.

The herb–disease relationships were merged from several paths and consist of 11 854 reliable associations in the stringently selected set (FDR_BH < 0.05). We found that 4.35% of the herb-disease associations were inferred by using both MM symptoms and ingredients as the intermediates (Figure 2C). The same pattern was also observed for the full set (Supplementary Figure S4A), the loosely selected set (Supplementary Figure S4B). The two paths for connecting herb–disease relationships via symptoms or ingredients can more or less be analogized to PDD and TDD. The SymMap data reveal that phenotype information facilitated the discovery of ethnopharmacological candidates, which will provide a valuable resource for translational medicine studies. Furthermore, we found that 35.34% (P-value < 0.05) and 31.95% (FDR_BH < 0.05) of herb–disease associations from manual curation can also be inferred from statistical inference (Supplementary Table S5), which further demonstrated the reliability of the statistical methods used in SymMap.

Finally, we integrated all six components of SymMap, as well as their pairwise relationships, including both direct and indirect associations, with the latter chosen from the stringently selected set with a FDR (BH) smaller than 0.05. We thus built a heterogeneous network including 32 281 nodes and 403 318 edges, with 106 721 edges representing direct associations and 296 597 edges representing indirect associations. The distributions of node degrees for the direct and indirect associations are quite similar (Figure 2D). Most nodes have a degree lower than 20, with a ratio of 95.98% for direct associations and 79.02% for indirect associations, which shows that the network is sparse in most parts. Furthermore, we analyzed the shared molecular interactions on disease–symptom associations using a previously published method (6). Consistent with previous observations, we found that when diseases are more similar in term of shared symptoms, they tended to be linked with each other through the underlying genes in their PPI network (Supplementary Figure S5). It further demonstrates the value of SymMap in connecting external symptom mapping and internal molecular mechanisms.

Functionality of SymMap

Users can browse, search and download the six components and their pairwise relationships through the SymMap web interface (Figure 3A). Users can click the search button in the homepage, input a query term in the search page to execute the search. A different search box is provided for each of the components of SymMap, with multiple types of search keys provided. For example, to search for a specific MM symptom, three types of search keys are permitted, including the symptom name, the external ID in a widely accepted database, and multiple aliases that are collected from diverse databases for the convenience of the users. All types of allowable search keys are described under the search boxes and further explained in the download page. And users can download all search terms in key files provided by SymMap. Furthermore, users can select similar keys immediately after inputting query terms using the autocomplete search functionality included in SymMap.

After searching SymMap, matches for the input query terms are displayed in the lower part of the search page in a summary table with the SymMap ID as the first column. Users are encouraged to click the hyperlink on the SymMap ID for detailed information. In the details page, we provide descriptive information and relationships with other components using network visualizations and tables. Furthermore, a list of all items in each of the six components can be navigated in the browse page, and these lists are also downloadable in the website.

Using the SymMap database

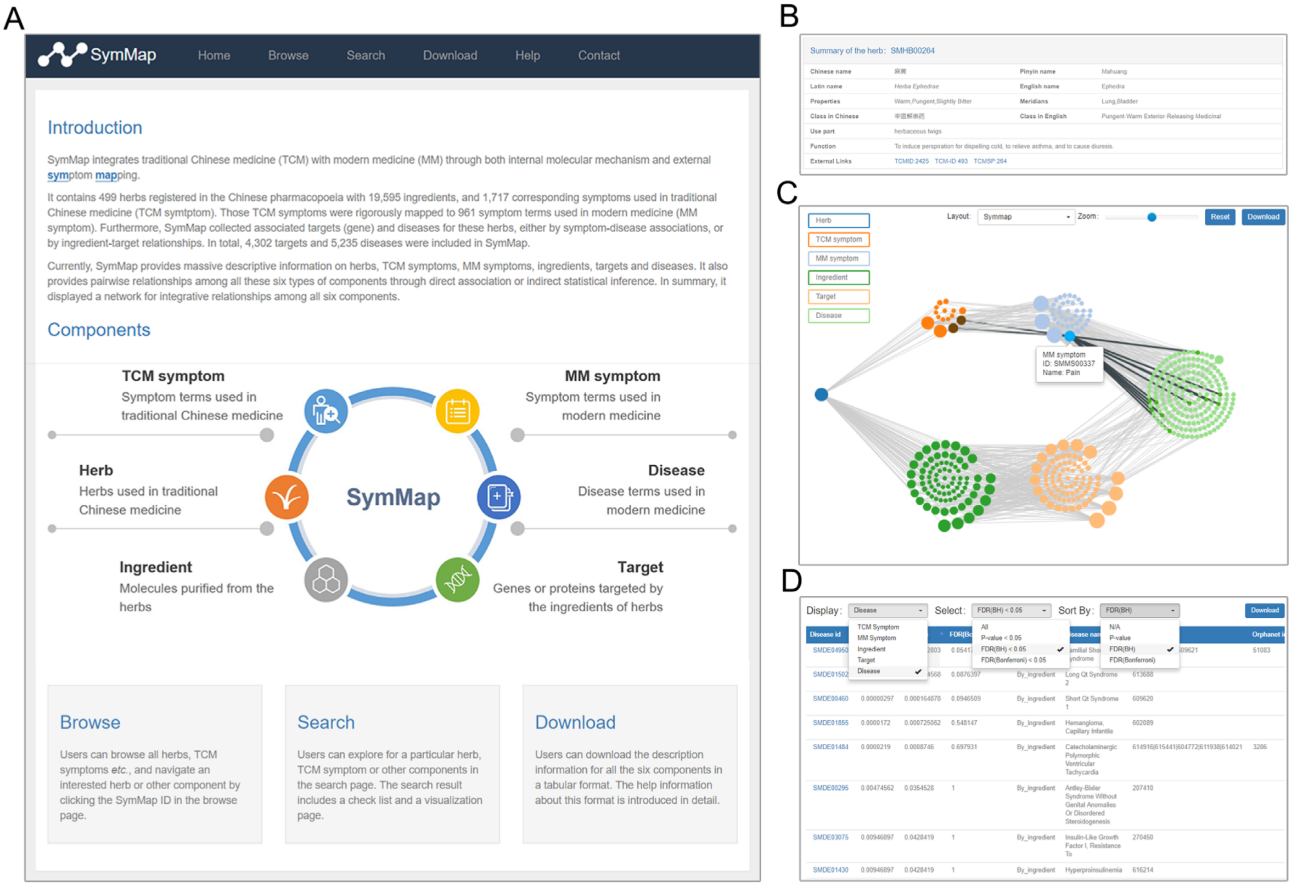
After browsing or searching SymMap, users can click the SymMap ID for each specific term to jump onto the details page, which provides a summary panel including descriptive information, a network panel visualizing the all-versus-all relationships among the six components, and a list panel showing tables of related items for the selected search key. The summary panel displays descriptive information for the search item (Figure 3B). In general, we provide three types of information: identification information (e.g. name and gene symbol), explanatory information (e.g. definition and class), and external IDs in other databases, which can be clicked directly to navigate into the database.

Next, the network panel provides a visualization of all related components for the search term (Figure 3C). The nodes in the network are colored and placed in different locations according to the source of the component. The node size is customized according to its degree in the network. When the user holds the mouse pointer over a node, the node will be enlarged, its related edges will be highlighted, and its ID and name will be shown in a balloon. In addition, each node in the picture can be hyperlinked to its own details page. We further provided control panels for users to change the network layout, to zoom in and out, and to download the network picture. To avoid the presence of an excessive number of nodes in a network, we chose the stringent selected set of indirect associations with FDR_BH <0.05 for the network visualization.

Finally, the list panel in the bottom of the detail page (Figure 3D) shows the information for the network visualization in tabular format, including five tables shown to represent five related components other than its own component. We provided three drop-down menus for users to customize the visualization. Firstly, the ‘display’ menu allows the users to select which relationship to access. Secondly, the ‘select’ menu enables the users to choose which dataset, including the full set and the subsets with different level of statistical inference, should be listed in the page. Thirdly, the ‘sort’ menu gives the users a capability to sort the items in the table according to SymMap IDs, *P*-values, FDRs (BH) and FDRs (Bonferroni). Furthermore, we added a ‘download’ button for users for bulk downloading.

DISCUSSION

A major goal of biomedical research is to elucidate phenotype–genotype relationships. Symptom phenotypes are diligently observed by physicians and are crucial for accurate clinical diagnosis and treatment. TCM symptoms have been utilized in clinical applications for millennia in a relatively large number of individuals. Therefore, TCM symptom–herb relationships provide tremendously valuable guidance for drug discovery programs. In this report, we present SymMap, a comprehensive database integrating TCM with MM via external symptom mapping and internal molecular mechanisms. The TCM symptoms in in SymMap, as well as their relationships with herbs and MM symptoms, were manually curated by a committee of 17 leading TCM experts. SymMap is the first publically available database containing comprehensive information regarding the relationships between TCM symptoms, TCM herbs and MM symptoms. Furthermore, users can access all-versus-all pairwise relationships between any two components in SymMap as direct associations obtained from database integration or indirect associations inferred based on statistical tests. Therefore, we have combined phenotype-based and target-based knowledge in SymMap to promote efficient phenotype-based compound screening under the guidance of current knowledge about targets for compounds and diseases. Users can easily access, navigate,



Introduction

SymMap integrates traditional Chinese medicine (TCM) with modern medicine (MM) through both internal molecular mechanism and external symptom mapping.

It contains 499 herbs registered in the Chinese pharmacopoeia with 19,595 ingredients, and 1,717 corresponding symptoms used in traditional Chinese medicine (TCM symptom). Those TCM symptoms were rigorously mapped to 961 symptom terms used in modern medicine (MM symptom). Furthermore, SymMap collected associated targets (gene) and diseases for these herbs, either by symptom-disease associations, or by ingredient-target relationships. In total, 4,302 targets and 5,235 diseases were included in SymMap.

Currently, SymMap provides massive descriptive information on herbs, TCM symptoms, MM symptoms, ingredients, targets and diseases. It also provides pairwise relationships among all these six types of components through direct association or indirect statistical inference. In summary, it displayed a network for integrative relationships among all six components.

Components

TCM symptom: Symptom terms used in traditional Chinese medicine

MM symptom: Symptom terms used in modern medicine

Herb: Herbs used in traditional Chinese medicine

Disease: Disease terms used in modern medicine

Ingredient: Molecules purified from the herbs

Target: Genes or proteins targeted by the ingredients of herbs

Browse: Users can browse all herbs, TCM symptoms *etc.*, and navigate an interested herb or other component by clicking the SymMap ID in the browse page.

Search: Users can explore for a particular herb, TCM symptom or other components in the search page. The search result includes a check list and a visualization page.

Download: Users can download the description information for all the six components in a tabular format. The help information about this format is introduced in detail.

Summary of the herb: SMHB00264

Chinese name	麻黄	Pinyin name	Mahuang
Latin name	*Herba Ephedrae*	English name	Ephedra
Properties	Warm,Pungent,Slightly Bitter	Meridians	Lung,Bladder
Class in Chinese	辛温解表药	Class in English	Pungent-Warm Exterior-Releasing Medicinal
Use part	herbaceous twigs		
Function	To induce perspiration for dispelling cold, to relieve asthma, and to cause diuresis.		
External Links	TCMID:2425 TCM-ID:493 TCMSP:264		

Layout: Symmap
Zoom:

MM symptom
ID: SMMS00337
Name: Pain

Display: Disease
Select: FDR(BH) < 0.05
Sort By: FDR(BH)

Disease id			FDR(Bo			isease nam		Orphanet i
SMDE04950		2803	0.05417			amilial Sho yndrome	09621	51083
SMDE01502		4568	0.0876397		By_ingredient	Long Qt Syndrome 2	613688	
SMDE00460	0.00000297	0.000164878	0.0946509		By_ingredient	Short Qt Syndrome 1	609620	
SMDE01855	0.0000172	0.000725062	0.548147		By_ingredient	Hemangioma, Capillary Infantile	602089	
SMDE01484	0.0000219	0.0008746	0.697931		By_ingredient	Catecholaminergic Polymorphic Ventricular Tachycardia	614916|615441|604772|611938|614021	3286
SMDE00295	0.00474562	0.0354528	1		By_ingredient	Antley-Bixler Syndrome Without Genital Anomalies Or Disordered Steroidogenesis	207410	
SMDE03075	0.00946897	0.0428419	1		By_ingredient	Insulin-Like Growth Factor I, Resistance To	270450	
SMDE01430	0.00946897	0.0428419	1		By_ingredient	Hyperproinsulinemia	616214	

Figure 3. An illustration of the SymMap search. (**A**) The index page of SymMap shows the database overview. (**B**) The summary panel in the details page shows descriptive information for the search item. (**C**) The network panel in the details page shows all related components for the search item, with nodes colored by their source component. Holding the mouse pointer over the node highlights the node and its related edges, while showing its ID and name, as well as a link to its details page. (**D**) The list panel shown in tables. For each search item, five tables can be selected for the five other related components. For each table, three datasets can be selected by the users: the full set, the loosely selected set with *P*-values smaller than 0.05, and two stringently selected sets with FDRs (Bonferroni and BH) smaller than 0.05. All related components can be downloaded by pressing the button at the upper right.

and visualize these data, as well as the relationships between database components, using the website interface for the SymMap database. We plan to continue to add data to SymMap as additional information becomes available, as well as to improve the user experience at the SymMap website.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Kui Xu for help with the network visualization. We thank Dr Zhen Li and Dr Zhaoqing Ba for critical comments during manuscript preparation.

FUNDING

National Key Research and Development Program of China [2018YFC1313000, 2018YFC1313001, 2017YFC1703506]; National Natural Science Foundation for Young Scholars of China [31701141, 31701149, 31501066]; National Natural Science Foundation of China [91740113, 81522051]. Innovation Project for Institute of Computing Technology, CAS [20186060]. Funding for open access charge: National Natural Science Foundation of China.

Conflict of interest statement. None declared.

REFERENCES

1. Swinney,D.C. (2013) Phenotypic vs. target-based drug discovery for first-in-class medicines. *Clin. Pharmacol. Ther.*, **93**, 299–301.
2. Swinney,D.C. and Anthony,J. (2011) How were new medicines discovered? *Nat. Rev. Drug Discov.*, **10**, 507–519.
3. Harvey,A.L., Edrada-Ebel,R. and Quinn,R.J. (2015) The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.*, **14**, 111–129.
4. Cragg,G.M. and Newman,D.J. (2013) Natural products: a continuing source of novel drug leads. *Biochim. Biophys. Acta*, **1830**, 3670–3695.
5. Qiu,J. (2007) Traditional medicine: a culture in the balance. *Nature*, **448**, 126–128.
6. Zhou,X., Menche,J., Barabasi,A.L. and Sharma,A. (2014) Human symptoms-disease network. *Nat. Commun.*, **5**, 4212.
7. Chen,K.K. (2012) A pharmacognostic and chemical study of ma huang (Ephedra vulgaris var. helvetica). 1925. *J. Am. Pharmacists Assoc.: JAPhA*, **52**, 406–412.

8. Tu,Y. (2011) The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine. *Nat. Med.*, **17**, 1217–1220.
9. 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.
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. 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.
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. 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.
14. Zhou,X., Wu,Z., Yin,A., Wu,L., Fan,W. and Zhang,R. (2004) Ontology development for unified traditional Chinese medical language system. *Artif. Intell. Med.*, **32**, 15–27.
15. Minguet,F., Salgado,T.M., Boogerd,L.V.D. and Fernandez-Llimos,F. (2015) Quality of pharmacy-specific Medical Subject Headings (MeSH) assignment in pharmacy journals indexed in MEDLINE. *Res. Soc. Admin. Pharm.*, **11**, 686–695.
16. Kuhn,M., Letunic,I., Jensen,L.J. and Bork,P. (2016) The SIDER database of drugs and side effects. *Nucleic Acids Res.*, **44**, D1075–D1079.
17. Nadkarni,P., Chen,R. and Brandt,C. (2001) UMLS concept indexing for production databases. *J. Am. Med. Informatics Assoc. Jamia*, **8**, 512.
18. Köhler,S., Vasilevsky,N.A., Engelstad,M., Foster,E., Mcmurry,J., Aymé,S., Baynam,G., Bello,S.M., Boerkoel,C.F. and Boycott,K.M. (2017) The Human Phenotype Ontology in 2017. *Nucleic Acids Res.*, **45**, D865–D876.
19. Wishart,D.S., Knox,C., Guo,A.C., Shrivastava,S., Hassanali,M., Stothard,P., Chang,Z. and Woolsey,J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.
20. Brown,G.R., Hem,V., Katz,K.S., Ovetsky,M., Wallin,C., Ermolaeva,O., Tolstoy,I., Tatusova,T., Pruitt,K.D., Maglott,D.R. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–D42.
21. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and Mckusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, 514–517.
22. Rath,A., Olry,A., Dhombres,F., Brandt,M.M., Urbero,B. and Ayme,S. (2012) Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum. Mutat.*, **33**, 803–808.
23. Fisher,R.A. (1922) On the interpretation of χ 2from contingency tables, and the calculation of P. *J. R. Statist. Soc. Ser. A*, **85**, 87–94.
24. Guney,E., Menche,J., Vidal,M. and Barábasi,A.L. (2016) Network-based in silico drug efficacy screening. *Nat. Commun.*, **7**, 10331.
25. Bland,J.M. and Altman,D.G. (1995) Multiple significance tests: the Bonferroni method. *BMJ*, **310**, 170.
26. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.. Ser. B (Methodological)*, **57**, 289–300.