BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 23 2014, pages 3379–3386
doi:10.1093/bioinformatics/btu560

Databases and ontologies

Advance Access publication August 24, 2014

ExpTreeDB: Web-based query and visualization of manually annotated gene expression profiling experiments of human and mouse from GEO

Ming Ni[1,†], Fuqiang Ye[1,†], Juanjuan Zhu[1,2], Zongwei Li[1], Shuai Yang[1], Bite Yang[1], Lu Han[1], Yongge Wu[2], Ying Chen[1], Fei Li[1,*], Shengqi Wang[1,3,*] and Xiaochen Bo[1,*]

[1]Beijing Institute of Radiation Medicine, Beijing 100850, [2]College of Life Sciences, Jilin University, Changchun 130012 and [3]Henan University of Traditional Chinese Medicine, Zhengzhou 450008, China

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Numerous public microarray datasets are valuable resources for the scientific communities. Several online tools have made great steps to use these data by querying related datasets with users' own gene signatures or expression profiles. However, dataset annotation and result exhibition still need to be improved.

Results: ExpTreeDB is a database that allows for queries on human and mouse microarray experiments from Gene Expression Omnibus with gene signatures or profiles. Compared with similar applications, ExpTreeDB pays more attention to dataset annotations and result visualization. We introduced a multiple-level annotation system to depict and organize original experiments. For example, a tamoxifen-treated cell line experiment is hierarchically annotated as 'agent→drug→estro gen receptor antagonist→tamoxifen'. Consequently, retrieved results are exhibited by an interactive tree-structured graphics, which provide an overview for related experiments and might enlighten users on key items of interest.

Availability and implementation: The database is freely available at http://biotech.bmi.ac.cn/ExpTreeDB. Web site is implemented in Perl, PHP, R, MySQL and Apache.

Contact: boxc@bmi.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 28, 2014; revised on July 31, 2014; accepted on August 15, 2014

1 INTRODUCTION

Comparing gene expression profiles of cultured cells, tissues or other samples at different states or time points is a common strategy to search for key factors or underlying mechanisms in biological and medical studies. Over the past decades, commercial microarray platforms have been used for global gene expression profiling, and numerous profiling datasets have been produced (Hoheisel, 2006). Gene Expression Omnibus (GEO), the public repository collecting microarray datasets (Edgar *et al.*, 2002), has included 32 856 series datasets across 1591 organisms produced by microarrays (January 24, 2014). These public datasets are valuable resources for biological, medical and pharmaceutical studies.

To use large numbers of public gene expression datasets, a useful strategy is to connect datasets or experiments with similar or opposite gene expression signatures. One important application of this strategy is to predict novel drug repositioning candidates or promising bioactive small molecules. The connectivity map (cMap) project has generated a large number of gene expression profiles of human tumor cell lines treated with over thousands of bioactive small molecules and systematically discovered potential connections among molecules, genes and diseases (Lamb, 2007; Lamb *et al.*, 2006). On using the datasets of the cMap project, Iorio *et al.* found 41 047 connections among 1302 drugs based on similarities of transcriptional responses following drug treatments; several potentially 'repositioned' drugs were predicted and verified, i.e. Topoisomerase inhibitors like *SN-38* and *Doxorubicin* were repositioned to functional inhibition of cyclin-dependent kinase 2 (Iorio *et al.*, 2010). Using similar strategy, Butte and his colleagues investigated the connections of global transcriptional responses between diseases and 164 small-molecule drug compounds; using rodent models, they predicted and validated that topiramate, an anticonvulsant drug, might be effective in inflammatory bowel disease (IBD), and cimetidine, an antiulcer drug, might be a previously undescribed therapeutic for lung adenocarcinoma (Dudley *et al.*, 2011; Sirota *et al.*, 2011).

Several algorithms have been developed to quantitatively compare expression profiles of global transcriptional responses (Lamb *et al.*, 2003; Plaisier *et al.*, 2010; Subramanian *et al.*, 2005). Among these, gene set enrichment analysis (GSEA) based on the Kolmogorov–Smirnov statistic (Hollander and Wolfe, 1973) is a widely used method (Subramanian *et al.*, 2005, 2007). Instead of individual genes, GSEA focuses on gene sets sharing biological functions or other information like chromosomal location and regulation. Specifically, GSEA compares a given gene set with a ranked gene list, and provides scores and statistics to show how the gene set is enriched at extremes (top or bottom) of the gene list.

Facilitated by these algorithms, various online tools and resources have emerged. The cMap, as we described previously, allows users to access and query with an algorithm later formalized in GSEA. Molecular Signatures Database (MSigDB) 3.0 has included >6700 well-annotated gene sets from human and

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

other model organisms and supports keyword search and GSEA-based enrichment analysis of the users' own gene sets (Liberzon *et al.*, 2011; Subramanian *et al.*, 2005). Unlike MSigDB used for querying annotated gene sets, online applications such as Microarray Rank Query (MARQ) and Gene Expression data Mining Toward Relevant Network Discovery (GEM-TREND) allow users to discover experiments in GEO that induce similar or opposite gene expression patterns to their own experiments with GSEA approach (Feng *et al.*, 2009; Vazquez *et al.*, 2010). Yi *et al.* (2007) reported a bioinformatics strategy named EXpression signature AnaLysis Tool (EXALT) to encode and compare gene expression signatures. Then, Wu *et al.* (2009) developed an online version of EXALT that includes >28 000 gene expression signatures derived from GEO and allows keywords and signature search. Other Web servers for analysis and annotation of GEO datasets like GeneChaser and GEOGLE also support multiple gene search (Chen *et al.*, 2008; Yu *et al.*, 2009). Besides these online applications, our group previously developed an R package named GeneExpressionSignature (Li *et al.*, 2013), which integrated the GSEA algorithm and could be easily used in users' own bioinformatics analysis pipelines.

To search among a large number of gene expression datasets, clear and comprehensive annotations of the original experiments and vivid exhibitions of retrieved results are important. Current tools usually display retrieved datasets in a long table list sorted by similarity scores or significance statistics. Although information such as experimental design and samples used is given, the results are not well organized and are difficult to browse or find items of interest. Moreover, to provide a biologically meaningful interpretation of retrieved results, selection of control groups and their corresponding experimental groups is critical. In this article, we present an online database ExpTreeDB, which includes manually curated gene expression datasets derived from experiments deposited in GEO. ExpTreeDB supports users' query in terms of either gene sets or global gene profiles based on a GSEA approach. We manually selected control and experimental groups for original GEO dataset records and implemented a multiple-level annotation system for these datasets according to original experimental treatments and description. In addition, users' query results are organized and visualized by interactive tree-structured graphics. Users can conveniently collapse or expand tree branches to focus on experiments of interest. ExpTreeDB includes 1351 reference ranked gene lists representing various experimental treatments in human and mouse. The database is available at http://biotech.bmi.ac.cn/ExpTreeDB.

2 METHODS

2.1 Gene expression profiling data collection

We downloaded global gene expression profiling datasets of *Homo sapiens* and *Mus musculus* from GEO. We selected GEO datasets (GDS) records produced by four platforms: Affymetrix mouse genome 430 2.0 array (GPL1261), Affymetrix mouse expression 430A array (GPL339), Affymetrix human genome U133 plus 2.0 array (GPL570) and Affymetrix human genome U133A array (GPL96). A total of 425 mouse and 511 human GDS records were obtained, which accounted for 43.69 and 44.86% of mouse and human GDS records (Mar 5, 2013), respectively.

2.2 Generating reference ranked gene lists

Reference ranked gene lists (RRGLs) were derived from GDS records. Every RRGL is composed of ranked upregulated and downregulated genes that represent global transcriptional response induced by certain experimental treatment or abnormal biological state. To generate RRGLs, we first identified control and experimental groups within GDS records. For a clear definition of perturbations and a better interpretation of transcriptional responses, we selected only GDS records with 'blank' control groups whose experimental subjects were at normal states or without treatments, such as adjacent-tumor tissue or samples treated with blank vehicle.

The workflow to collect information and filter GDS records is shown in Figure 1. The keywords for determining a blank control group and the GDS records excluded are provided in Supplementary Table S1 and S2, respectively. For experimental groups, one or multiple perturbations may be imposed on experimental subjects, such as agent treatment or gene knockdown at disease state.

Next, blank control groups were matched to experimental groups within the same GDS record. Every experimental group and its corresponding control group were manually matched to generate a RRGL. Home-made PERL scripts and R package GeneExpressionSignature that we previously developed were used to produce RRGLs (Li *et al.*, 2013). In brief, for a given experimental–control group pair, we calculated gene expression fold-change lists among all possible sample pairs. Then, the fold-change lists derived from different pairs were merged by Spearman's Footrule (Diaconis and Graham, 1977) and Borda Merging Method (Lin, 2010). The final RRGL was a one-column ranked gene list sorted by fold changes, representing cell response to a specific experimental treatment under a certain condition. We compared RRGLs produced by this method with a student's *t*-test approach. We found a median Pearson coefficient correlation value of 0.88 and a mean value of 0.84 (data not shown).

2.3 Multiple-level annotation of RRGLs

To annotate RRGLs, a four-level annotation system was manually implemented based on original experimental information. Basically, the detailed treatments on experimental subjects or disease states were defined as the last-level annotations, which were classified into five top-level

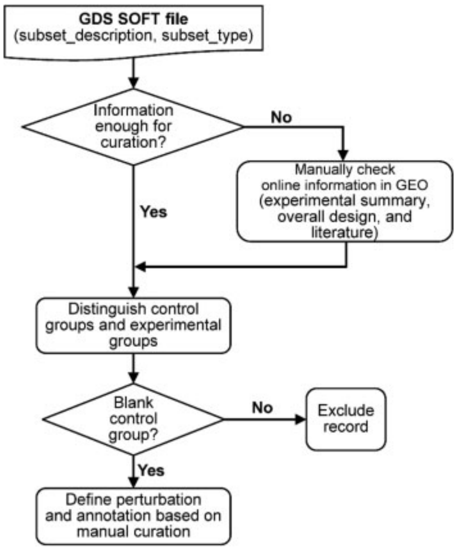
GDS SOFT file
(subset_description, subset_type)
Information enough for curation?
No
Manually check online information in GEO (experimental summary, overall design, and literature)
Yes
Distinguish control groups and experimental groups
Blank control group?
No
Exclude record
Yes
Define perturbation and annotation based on manual curation

Fig. 1. The workflow for selecting GDS records and collecting information

categories: ‘agent’, ‘stress’, ‘disease state’, ‘infection’ and ‘genotype variation’.

We should address that although GDS records have a ‘subset_type’ item in the experimental description, these somehow seem arbitrary. For example, ‘bacterial infection’ or ‘virus infection’ can be marked as ‘infection’, as well as ‘protocol’ and ‘agent’ in some cases (see GDS3298, GDS3034 and GDS1478); ‘cigarette smoking’ or ‘smoke’ is classified as both ‘stress’ and ‘agent’ (see GDS3709, GDS2990). In ExpTreeDB, we manually ensured uniqueness of experimental annotations. Additionally, ‘protocol’ subset type of human GDS records actually contains various experiment types. In ExpTreeDB, we replaced ‘protocol’ with clearer experiment annotations like ‘genotype variation’, ‘infection’ or ‘stress’ (see GDS3578, GDS3034 and GDS906).

The principles to add the second-level and third-level annotations within the five top categories are depicted below.

(1) Disease state. We referred to the World Health Organization International Classification of Diseases, 10th version (ICD-10). From the disease state description of GDS records or related literature, second-level annotations were defined as the ICD-10 chapters and third-level annotations as classification domain nearest to fourth-level annotation. For example, samples from individuals harboring papillary thyroid carcinoma were marked with ‘disease state→Neoplasms→Thyroid and other endocrine glands→papillary thyroid carcinoma’, where ‘Neoplasms’ is the top classification domain (ICD-10 chapter) covering ‘papillary thyroid carcinoma’, and ‘Thyroid and other endocrine glands’ is the lowest classification domain describing ‘papillary thyroid carcinoma’. For most RRGLs under ‘disease state’, we obtained their corresponding items of ICD-10, and the unclassified were labeled as ‘other’.

(2) Agent. If an agent could be referred to anatomical therapeutic chemical (ATC) classification systems, the corresponding RRGLs were labeled with ‘drug’ at second-level annotations. Otherwise, according to the agent types, the second-level annotations could be ‘small molecule’, ‘protein’, ‘lipid’, ‘diet’, ‘chemical’ or ‘other’. The third-level annotations were manually curated to represent the function or subtypes of agents, like ‘inflammation mediator’ and ‘cytokine’. For example, a RRGL of the effect of letrozole on breast cancer tumors could be annotated as ‘agent→drug→antiestrogen agent→letrozole’.

(3) Genotype variation. The last-level annotation was the object gene name. The second-level annotation indicated that the gene was protein coding or microRNAs. The third-level annotation recorded the experiment protocol or gene expression status, such as ‘knockout’, ‘knockin’, ‘transfection’, ‘overexpression’ or ‘induction’.

(4) Infection. Annotations were based on microorganisms. The second-level annotations of ‘infection’ RRGLs were microorganism domains: ‘bacteria’, ‘virus’, ‘fungi’ or ‘protozoan’. The third-level annotations were based on disease symptom or effect induced by pathogens and/or pathogen types. For example, RRGLs of *Aeromonas caviae* infection were annotated as ‘gram-negative, gastroenteritis-associated pathogen’ at third-level.

(5) Stress. We classified ‘stress’ into three categories, with ‘physical’, ‘psychological’ or ‘other’ as second-level annotations. Radiation, hypoxia and mechanical damage were typical third-level annotations of physical stress. We also classified cigarette smoking and exposure to pollutants into physical stress. Patient caregiver chronic stress was the only psychological stress annotation. Similar to ‘infection’, the third-level annotation depicted stress effect or subtype of stress, like ‘respiratory tract-harmful stimulus’ and ‘ionizing radiation’. A RRGL describing the chronic stress of patient caregiver could be labeled as ‘stress→psychological→patient caregiver-associated→chronic stress’.

The annotation of RRGLs allowed for multiple perturbations at all the four levels, such as a cell line under both genotype variation and agent treatment. For the four-level annotation of RRGLs, we only considered perturbations that might induce transcriptional responses. Other experimental information, such as tissue types, time points, individual gender, drug ATC codes and gene official symbols, would be provided as an additional field of RRGL annotations.

2.4 Querying related RRGLs (algorithm and parameters)

In case users upload raw gene expression profiling data, ExpTreeDB first obtains a gene signature from the uploaded data, and then performs query with the signature. Specifically, a RRGL is first generated from the user’s dataset, and ExpTreeDB extracts the top and bottom of RRGL as gene signature. We scanned the size of signatures, and found that a threshold of 500 up- or downregulated genes is suitable (Supplementary Material and Supplementary Fig. S1).

The R package GeneExpressionSignature, which integrates GSEA (Li *et al.*, 2013; Subramanian *et al.*, 2005), is implemented to compare user’s signature with RRGLs within ExpTreeDB. Enrichment scores for both upregulated genes (ES_{up}) and downregulated genes (ES_{down}) are obtained, and similarity is defined as $(ES_{up}-ES_{down})/2$. On condition that user uploads signature containing only upregulated genes or downregulated genes, similarity is either ES_{up} or ES_{down}.

ExpTreeDB will return the top 5% RRGLs ranked by enrichment scores on the condition that their *P*-values are <0.05. To derive *P*-values, ExpTreeDB calculates similarities between the users’ queries with 10 000 pre-generated random RRGLs, denoted by $\{S_r\}$. Given S_n as the similarity value between user’s query and a native RRGL, and n as the rank number of S_n in $\{S_r\}$, *P*-value is defined as $n/10\,001$.

2.5 Tree-structured visualization of related RRGLs

Based on the hierarchical annotations of gene expression experiment, ExpTreeDB generates an interactive tree visualization of related RRGLs. A JavaScript tool named dTree (dTree Version 2.05) is implemented to draw the tree. The root of the tree stands for the user’s query. The five top-level branches of tree are ‘disease state’, ‘agent’, ‘genotype variation’, ‘infection’ and ‘stress’, and sub-level branches and leaves are the sub-level annotations and the related RRGLs, respectively. Tree branches can be interactively collapsed and expanded. A table recording more detailed information of related RRGLs is exhibited below the tree, including the original GDS record number, the similarity score and rank, *P*-values and other experiment information.

2.6 Clustering and visualization of related RRGLs

To exhibit connections among returned related RRGLs, ExpTreeDB visualizes how the RRGLs are clustered by their differences. In details, pair-wise distances of all RRGLs of human or mouse was pre-generated. We used the MATLAB version of affinity propagation clustering algorithm proposed by Frey *et al.* (Frey and Dueck, 2007) to cluster RRGLs by distances. On the result Web page, ExpTreeDB presents only a network of related RRGLs in clusters, and the edges within the clusters are visible if the two RRGLs are top 5% similar among all possible human or mouse RRGL pairs.

2.7 Gene ontology analysis

For RRGLs within ‘genotype variation→coding gene’ category, we obtained the fourth-level annotations and manually transferred them into official gene symbols provided by the HUGO Gene Nomenclature Committee. The human and mouse gene symbol lists were uploaded to the Protein ANalysis THrough Evolutionary Relationships) classification system for Gene Ontology (GO) functional classification (Thomas *et al.*, 2003).

3 RESULTS

3.1 ExpTreeDB overview

ExpTreeDB consists of four parts: a submission page, a RRGL repository, a GSEA-centered calculation module and a result exhibition page (Fig. 2 and Supplementary Fig. S2). The RRGL repository is the base of ExpTreeDB. With blank control groups as baselines, RRGLs represent cell global transcriptional responses for various perturbations like treatments or abnormal states. These RRGLs are manually curated and hierarchically organized. Currently, ExpTreeDB contains 906 and 445 RRGLs in human and mouse, respectively.

A GSEA-centered calculation pipeline is used in ExpTreeDB, and users can query RRGL database to find similar gene expression signatures produced by certain experiments. ExpTreeDB enables queries from users by handling differentially expressed gene sets (gene signatures) or global gene expression profiles with assigned control and experimental groups. Based on the multiple-level annotations of RRGLs, ExpTreeDB will export related RRGLs to an interactive tree-structured graphic followed by a table containing detailed information. The matching details of query and a given related RRGL can be accessible in a new Web page. Furthermore, a network generated by clustering of related RRGLs is also displayed, exhibiting underlying connections among these RRGLs.

3.2 RRGL annotations and statistics

We implemented a hierarchical four-level annotation system on these RRGLs. For instance, an experiment consisting of a cell line treated with the drug tamoxifen would be annotated as 'agent→drug→estrogen receptor antagonist→tamoxifen'. Other experimental information like cell lines, time points, subject information, is also provided if any. The fourth-level annotations depict the specific perturbations inducing transcriptional responses, and the second- and third-level annotations were manually added based on the characteristics of each experiment type (see Methods). In sum, through the multiple-level annotation system, these RRGLs can be both well-organized and annotated.

The four-level annotations and their statistics of human and mouse RRGLs are illustrated in Figure 3 and Supplementary Figure S3, and the complete annotations can be downloaded on ExpTreeDB Web site (Letunic and Bork, 2007, 2011). Because the mouse is an important model animal, more than half of the mouse RRGLs (227, 59.5%) belong to the 'genotype variation' category. In human, only 9.7% of RRGLs (88 RRGLs) belong to that category. A total of 43 human and 140 mouse unique protein-coding genes are involved within these RRGLs. Functional classification based on GO shows that these human and mouse genes have similar distributions of GO term assignments (Fig. 4). Unlike the relatively high frequency of protein-coding genes, only five microRNAs are involved in human RRGLs (*miR-34*, *miR-155*, *miR-124*, *miR-335* and *miR-K12-11*) and 2 microRNAs in mouse (*miRNA-1-2* and *miR-290*) RRGLs.

The largest top-level category for human RRGLs is 'disease state' (270, 29.8%). For the second-level annotations, 'neoplasm' accounts for nearly half of these RRGLs (129, 47.8%), corresponding to 24 neoplasm types (Supplementary Table S3). 'Disease of the digestive system' and 'Diseases for the nervous system' are tied for the second place (8.1% of human 'disease state' RRGLs). On the other hand, we are interested in the various biological specimens used in these human disease studies and referred to the original 87 GDS records. We found that 56 GDS records involved human tissues as experimental subjects, and 20 GDS records used blood or bone marrow cells. Cultured cells were also used in 11 GDS records, such as lymphoblastoid cell lines derived from individuals (GDS2779) and pulmonary artery endothelial cells to study response to sickle cell plasma (GDS1257).

'Agent' was a major category for both human (446, 49.2%) and mouse (125, 28.1%) RRGLs, and exhibits great diversity, including drugs, small molecules, proteins/peptides, chemicals, biomaterials and others. RRGLs marked with 'drug' were of special interest for drug reposition research. A total of 33 human and 16 mouse GDS records involved 25 and 19 drugs, respectively. These drugs were correlated to 59 ATC codes, and 14 ATC codes were in common for human and mouse. 'Stress' and 'Infection' categories account for a small proportion of the whole RRGL repository. Nonetheless, some interesting and important perturbations like cigarette smoking, hypoxia and HIV infection belong to these two categories.

3.3 Query and export

Both gene signatures and raw gene expression profiles are supported by ExpTreeDB. For gene signatures, the input

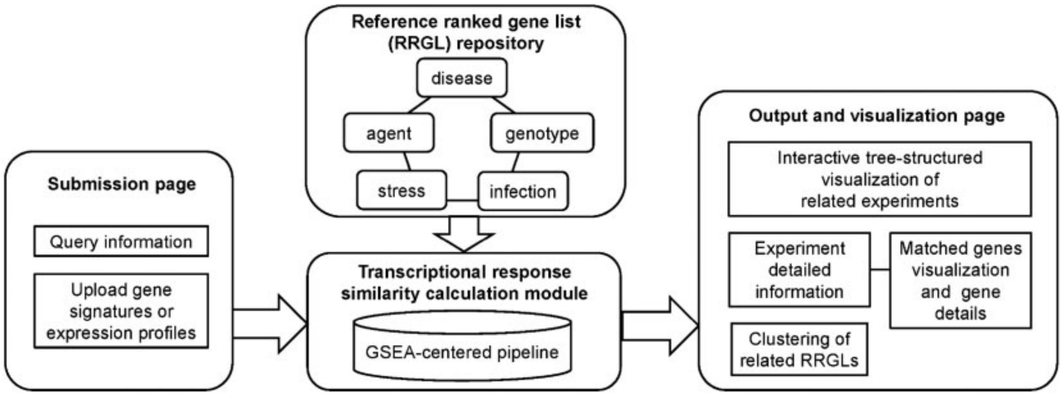
Reference ranked gene list
(RRGL) repository
disease
agent
genotype
stress
infection
Submission page
Query information
Upload gene signatures or expression profiles
Transcriptional response similarity calculation module
GSEA-centered pipeline
Output and visualization page
Interactive tree-structured visualization of related experiments
Experiment detailed information
Matched genes visualization and gene details
Clustering of related RRGLs

Fig. 2. The four parts of ExpTreeDB structure and the interfaces for data submission and result exhibition

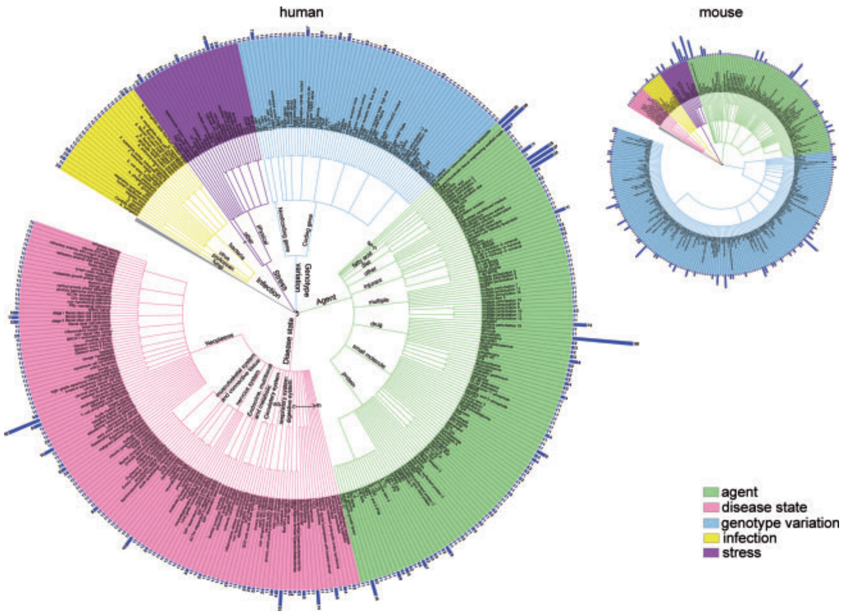

Fig. 3. Circular hierarchical tree illustrations of human and mouse RRGL annotations generated by using the Interactive Tree of Life (iTOL) online application (Letunic and Bork, 2007, 2011). For human tree illustration (on the left side), the fourth-level annotations are represented as leaves of the trees and the other levels as tree branch nodes. The third-level annotations are omitted because of space limits. The blue bars surrounding the tree denote the total number of human RRGLs corresponding to each four-level annotation. The tree branch denoting multiple perturbations at first level is collapsed (gray branch). Some annotations are replaced by identifiers (see Supplementary Material). For space limits, a small-sized tree illustration of mouse RRGL annotations (on the right) is shown for comparison (see full-size illustration in Supplementary Fig. S3)

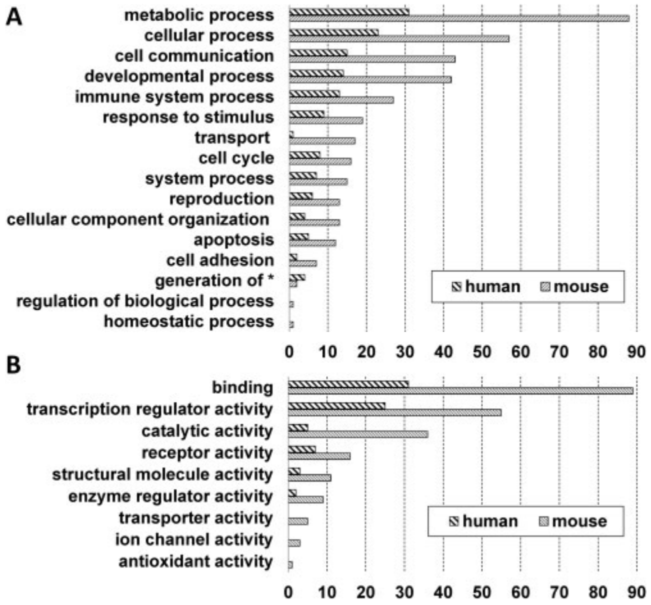

Fig. 4. Functional classification of human and mouse genes based on GO assignments. Hit numbers of (**A**) GO biological process terms and (**B**) GO molecular function terms are displayed. Hybrid genes were excluded from GO analysis. The term marked with a star is short for 'generation of precursor metabolites and energy'

identifier is flexible (Entrez Gene ID, Ensembl ID, Genbank accession number, gene symbol and Affymetrix Probe ID). Upregulated genes and/or downregulated genes should be uploaded with a constrained range of gene numbers (10~500 for Affymetrix Probe ID, and 5~300 for other gene identifiers) when users submit their signatures. Alternatively, users can directly upload gene expression profiles. The control and experimental group for these profiles have to be defined for ExpTreeDB to generate a gene signature for further calculations. ExpTreeDB supports datasets produced by a number of Affymetrix platforms or next-generation sequencing technology (Supplementary Table S4). A plain tab-separated gene expression value matrix with column names as sample identifiers and row names as gene identifiers is also allowed, which can be easily transformed from other gene expression profile formats.

ExpTreeDB converts all other gene identifiers into Affymetrix Probe IDs for following calculation. The GSEA-centered pipeline is used to calculate pair-wise enrichment scores and statistical significance between the user's dataset and RRGLs within ExpTreeDB. RRGLs with enrichment scores ranking top 5% and P-values <0.05 would be retrieved for output.

The result page is mainly composed of three parts: an interactive tree structure of annotations of related RRGLs, a table of detailed RRGL information and a clustering network of related RRGLs (Supplementary Fig. S2). The tree is constructed based on the multiple-level annotation system of RRGLs, with branch nodes as sub-category annotations and leaves as RRGL items held in the database (see example in Fig. 5). From the tree, users can obtain a global view and classifications of how their submitted data connected with previous reported experiments. To focus on their topic of interest, users can expand and collapse tree nodes at any level, or choose to display a particular part of the tree. For example, to mine potential drug repositions, user can expand the 'agent→drug' node on the tree, and collapse the others. The effect of retrieved drugs and drug names will be displayed under the ‘agent→drug’ node.

To obtain the detailed information of retrieved RRGLs, RRGL accession numbers on the tree are linked to corresponding rows of the table below the tree. The table contains original experiment description, GDS record accession numbers with hyperlinks to GEO Web site, GSEA-centered calculation results (rank, enrichment score, *P*-value), four-level annotations and other supporting information like cell line, drug ATC code or time points. We also provide a page to illustrate the matching details between query and a given related-RRGL. The positions of query genes within the RRGL are visualized by a colored bar, and gene details such as GO terms are presented in a table (see Supplementary Fig. S2). To determine connections among these RRGLs, ExpTreeDB pre-generated a network by clustering enrichment score matrix of all RRGL pairs. The connections among retrieved RRGLs will be displayed. The output page could be retrieved by a task identifier within three months.

3.4 Sample case: to mine for perturbations similar to hypoxia

We downloaded gene expression datasets of human renal proximal tubule epithelial cells subjected to hypoxia (GDS3524, GSE12792) (Beyer *et al.*, 2008). To mine perturbations inducing transcriptional responses similar to hypoxia, we submitted the SOFT format datasets (GSE12792_family.soft.gz) to ExpTreeDB. Samples 4, 5 and 6 within the profile series were assigned as the experimental group and samples 1, 2 and 3 as control group, allowing comparison of the cell response under hypoxic condition (at 1% oxygen for 24 h) with those under normal oxygen content. The calculation process took 43 s, and results could be accessed with task identifier 463. A total of 46 related RRGLs were returned.

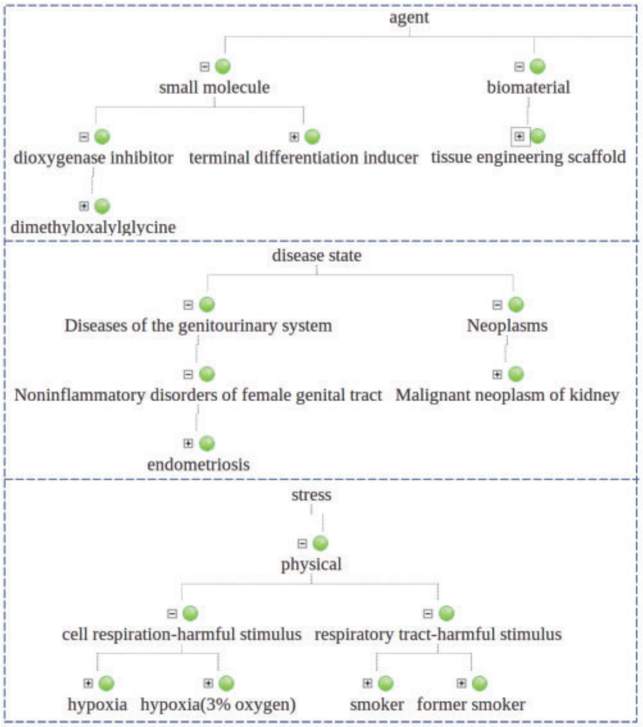
agent
small molecule
biomaterial
dioxygenase inhibitor
terminal differentiation inducer
tissue engineering scaffold
dimethyloxalylglycine
disease state
Diseases of the genitourinary system
Neoplasms
Noninflammatory disorders of female genital tract
Malignant neoplasm of kidney
endometriosis
stress
physical
cell respiration-harmful stimulus
respiratory tract-harmful stimulus
hypoxia
hypoxia(3% oxygen)
smoker
former smoker

Fig. 5. The tree-structured visualization of related RRGLs for hypoxia-induced cell responses. For space limits, the tree branches of ‘agent’, ‘disease state’ and ‘stress’ are, respectively, shown. The full tree can be accessed through task identifier 463

From the annotation tree of the returned RRGLs, we obtained an overview of related experiment types and specific treatments (Fig. 5). These RRGLs were classified into three top-level categories: ‘stress’, ‘disease state’ and ‘agent’. The subclassifications within ‘stress’ and ‘disease state’ were simple. Besides the hypoxia’s effect on various cell lines, cigarette smoking, endometriosis and renal cell carcinoma (RCC) at different stages were three perturbation types that induced cell responses similar to the query. Cigarette smoking is known to decrease tissue or blood flow oxygen (Fan *et al.*, 2012; Hoff *et al.*, 2012; Jensen *et al.*, 1991; Sorensen *et al.*, 2009). The query results implied that the state of the non-diseased lung tissues from both ‘smoker’ and ‘former smoker’ is related to hypoxia, with ‘smoker’ at a higher rank. A major feature for solid tumors is hypoxia (Brahimi-Horn *et al.*, 2007; Brown and Wilson, 2004; Semenza, 2007; Vaupel and Mayer, 2007), and RCC has been reported to overexpress hypoxia-inducible factor-1α (*HIF-1α*) when compared with benign tissue (Klatte *et al.*, 2007). However, no RRGLs corresponding to other solid tumor besides RCC were returned. As most RCC RRGLs ranked at the bottom of the 46 RRGLs, other solid tumor RRGLs might be filtered by enrichment score rank or significance threshold. The connection of endometriosis with hypoxia was interesting. A study in mice suggested that hypoxia appears to be important for endometriosis and results in upregulation of *HIF-1α* (Becker *et al.*, 2008).

Moreover, the RRGLs in the ‘agent’ category exhibited great diversity in terms of agent types. From the annotation tree, we found three types of drugs, two types of small molecules, cytokine, carcinogen (nickel chloride) and biomaterial. At the fourth-level annotations, RRGLs annotated as ‘dimethyloxalylglycine’ and ‘collagen/chondroitin tissue engineering scaffold mesh’ were top-ranked among the retrieved RRGLs. The literature associated with these also reported that these two types of agents induced gene expression patterns like hypoxia (Elvidge *et al.*, 2006; Klapperich and Bertozzi, 2004). The validation of the relationship between hypoxia and other agent treatments, such as the growth inhibitor imatinib and a small molecule that induces differentiation, might need further investigation.

Another five case studies are included in Supplementary Material, which are provided to exhibit reliability of ExpTreeDB outputs and advantage of the multiple-annotation system in user experience.

4 DISCUSSION

Currently, many other online applications querying gene signature databases have been developed, such as MSigDB, MARQ, GEM-TREND, EXALT and GeneChaser (Chen *et al.*, 2008; Feng *et al.*, 2009; Liberzon *et al.*, 2011; Vazquez *et al.*, 2010; Wu *et al.*, 2009), and provide users with great convenience. A tendency of these types of applications is to cover more and more datasets (Liberzon *et al.*, 2011), and algorithms to quantify expression profile or signature similarities have also been greatly emphasized.

ExpTreeDB resembles these tools in many aspects such as data sources and enrichment score calculation algorithm. Vazquez *et al*. comprehensively summarized and compared the features of various online applications, which could be referred in (Vazquez *et al*., 2010). Compared with these applications, ExpTreeDB exhibits a disadvantage in organism coverage (human and mouse) and data content (906 human and 445 mouse RRGLs). However, previous studies did not provide a well-depicted standard for control group or reference dataset selection, which might make it difficult to interpret the perturbation inducing the cell response (see example in Supplementary Material). Raw gene expression profiles are also supported by ExpTreeDB, and the threshold of 500 most upregulated and downregulated genes was chosen to generate signatures from these data. Though this threshold might not be suitable for datasets with very few differentially expressed genes, it could be applied to other datasets in most cases. Moreover, previous applications generally exported retrieved results in a table list sorted by relevance with users' query, which might be inconvenient for browsing or finding items of interest. MSigDB classified gene sets into seven major collections and several sub-collections according to genes' derivation, including gene locations in chromosomes, pathway genes and motif genes, which is an approach essentially different from the experiment-based classification in ExpTreeDB.

In ExpTreeDB, GDS records were manually examined to generate RRGLs. Datasets without blank control group were excluded. The stringent requirement for control groups remarkably decreased GDS records included in ExpTreeDB, but ensured a clear biological interpretation of perturbations inducing cell transcriptional responses. On top of that, we implemented a multiple-level annotation system for original experiments, which greatly differed from other similar online applications. The annotations briefly and comprehensively depicted experiment classifications and designs. By this approach, data within ExpTreeDB were well organized, and query results could be vividly visualized by an interactive and informative tree structure. Therefore, the multiple-level annotation system could greatly improve users' experience in result interpretation.

In all, ExpTreeDB is an online database for querying gene expression datasets of human and mouse derived from GEO. Users can upload global gene profiling data or gene signatures to mine related previously reported experiments. ExpTreeDB integrates a widely used algorithm and public gene expression datasets, but characterizes itself by manually annotated experiments and tree visualization of retrieved results. We also have a plan to update the database repository by including GEO series records in the next version of ExpTreeDB.

Funding: This work was supported by National Natural Science Foundation of China (Grant No. 81273488, 81230089), National Key Technologies R&D Program for New Drugs (Grant No. 2012ZX09301-003). Program of International S&T Cooperation (Grant No. 2014DFB30020).

Conflict of interest: none declared.

REFERENCES

Becker,C.M. *et al*. (2008) 2-methoxyestradiol inhibits hypoxia-inducible factor-1{alpha} and suppresses growth of lesions in a mouse model of endometriosis. *Am. J. Pathol*., **172**, 534–544.

Beyer,S. *et al*. (2008) The histone demethylases JMJD1A and JMJD2B are transcriptional targets of hypoxia-inducible factor HIF. *J. Biol. Chem*., **283**, 36542–36552.

Brahimi-Horn,M.C. *et al*. (2007) Hypoxia and cancer. *J. Mol. Med. (Berl)*, **85**, 1301–1307.

Brown,J.M. and Wilson,W.R. (2004) Exploiting tumour hypoxia in cancer treatment. *Nat. Rev. Cancer*, **4**, 437–447.

Chen,R. *et al*. (2008) GeneChaser: identifying all biological and clinical conditions in which genes of interest are differentially expressed. *BMC Bioinformatics*, **9**, 548.

Diaconis,P. and Graham,R. (1977) Spearman's footrule as a measure of disarray. *J. R. Stat. Soc*., **39**, 262–268.

Dudley,J.T. *et al*. (2011) Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med*., **3,** 96ra76.

Edgar,R. *et al*. (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*., **30**, 207–210.

Elvidge,G.P. *et al*. (2006) Concordant regulation of gene expression by hypoxia and 2-oxoglutarate-dependent dioxygenase inhibition: the role of HIF-1alpha, HIF-2alpha, and other pathways. *J. Biol. Chem*., **281**, 15215–15226.

Fan,G.B. *et al*. (2012) Changes of oxygen content in facial skin before and after cigarette smoking. *Skin Res. Technol*., **18**, 511–515.

Feng,C. *et al*. (2009) GEM-TREND: a web tool for gene expression data mining toward relevant network discovery. *BMC Genomics*, **10**, 411.

Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.

Hoff,C.M. *et al*. (2012) Effect of smoking on oxygen delivery and outcome in patients treated with radiotherapy for head and neck squamous cell carcinoma—a prospective study. *Radiother Oncol*., **103**, 38–44.

Hoheisel,J.D. (2006) Microarray technology: beyond transcript profiling and genotype analysis. *Nat. Rev. Genet*., **7**, 200–210.

Hollander,M. and Wolfe,D.A. (1973) *Nonparametric Statistical Methods*. Wiley, New York.

Iorio,F. *et al*. (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl Acad. Sci. USA*, **107**, 14621–14626.

Jensen,J.A. *et al*. (1991) Cigarette smoking decreases tissue oxygen. *Arch. Surg*., **126**, 1131–1134.

Klapperich,C.M. and Bertozzi,C.R. (2004) Global gene expression of cells attached to a tissue engineering scaffold. *Biomaterials*, **25**, 5631–5641.

Klatte,T. *et al*. (2007) Hypoxia-inducible factor 1 alpha in clear cell renal cell carcinoma. *Clin. Cancer Res*., **13**, 7388–7393.

Lamb,J. (2007) The connectivity map: a new tool for biomedical research. *Nat. Rev. Cancer*, **7**, 54–60.

Lamb,J. *et al*. (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Lamb,J. *et al*. (2003) A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell*, **114**, 323–334.

Letunic,I. and Bork,P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127–128.

Letunic,I. and Bork,P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res*., **39**, W475–W478.

Li,F. *et al*. (2013) GeneExpressionSignature: an R package for discovering functional connections using gene expression signatures. *Omics*, **17**, 116–118.

Liberzon,A. *et al*. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

Lin,S. (2010) Space oriented rank-based data integration. *Stat. Appl. Genet. Mol. Biol*., **9,** Article20.

Plaisier,S.B. *et al*. (2010) Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Res*, **38**, e169.

Semenza,G.L. (2007) Hypoxia and cancer. *Cancer Metastasis Rev*., **26**, 223–224.

Sirota,M. *et al*. (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med*., **3,** 96ra77.

Sorensen,L.T. *et al*. (2009) Acute effects of nicotine and smoking on blood flow, tissue oxygen, and aerobe metabolism of the skin and subcutis. *J. Surg. Res*., **152**, 224–230.

Subramanian,A. *et al*. (2007) GSEA-P: a desktop application for gene set enrichment analysis. *Bioinformatics*, **23**, 3251–3253.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Thomas,P.D. *et al*. (2003) PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res*., **31**, 334–341.

Vaupel,P. and Mayer,A. (2007) Hypoxia in cancer: significance and impact on clinical outcome. *Cancer Metastasis Rev*., **26**, 225–239.

Vazquez,M. *et al*. (2010) MARQ: an online tool to mine GEO for experiments with similar or opposite gene expression signatures. *Nucleic Acids Res*., **38**, W228–W232.

Wu,J. *et al*. (2009) Web-based interrogation of gene expression signatures using EXALT. *BMC Bioinformatics*, **10**, 420.

Yi,Y. *et al*. (2007) Strategy for encoding and comparison of gene expression signatures. *Genome Biol*., **8**, R133.

Yu,Y. *et al*. (2009) GEOGLE: context mining tool for the correlation between gene expression and the phenotypic distinction. *BMC Bioinformatics*, **10**, 264.