Deep functional synthesis: a machine learning approach to gene functional enrichment

Sheng Wang[1,2,*], Jianzhu Ma[3,*], Samson Fong[3,4], Stefano Rensi[1,2],
Jiawei Han[5], Jian Peng[5], Dexter Pratt[3], Russ B. Altman[1,2#], Trey Ideker[3,4,#]

[1]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA.
[2]Department of Genetics, Stanford University, Stanford, CA 94305, USA.
[3]School of Medicine, University of California San Diego, San Diego, CA 92093, USA
[4]Department of Bioengineering, University of California San Diego, San Diego, CA 92093, USA
[5]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[*]These authors contributed equally to this work

[#]Email:russ.altman@stanford.edu, tideker@ucsd.edu

ABSTRACT

Gene functional enrichment is a mainstay of genomics, but it relies on manually curated databases of gene functions that are incomplete and unaware of the biological context. Here we present an alternative machine learning approach, Deep Functional Synthesis (DeepSyn), which moves beyond gene function databases to dynamically infer the functions of a gene set from its associated network of literature and data, conditioned on the disease and drug context of the current experiment. Using a knowledge graph with 3,048,803 associations between genes, diseases, drugs, and functions, DeepSyn obtained accurate performance (range 0.74 AUC to 0.96 AUC) on a variety of biological applications including drug target identification, gene set functional enrichment, and disease gene prediction.

Availability: The DeepSyn codebase is available on GitHub at http://github.com/wangshenguiuc/DeepSyn/ under an open source distribution license.

MAIN TEXT

A common outcome of genomic analysis is the discovery of gene sets underlying specific biological functions. For instance, gene expression analyses produce sets of genes that are differentially expressed across conditions, or that cluster by expression similarity. Proteomics experiments produce lists of proteins and, by implication, their encoding genes, and so on. In all of these cases, the basic hypothesis is that the identified genes work coherently towards the same biological processes or functions. To label these functions, one turns to functional enrichment analysis. Among the numerous approaches that have been developed (Rhee et al., 2008; Zhou et al., 2017), some of the more widely used ones are the hypergeometric statistic (Breitling et al., 2004; Huang et al., 2009; Pomaznoy et al., 2018; Zeeberg et al., 2003) and gene set enrichment analysis (GSEA) (Al-Shahrour et al., 2007; Backes et al., 2007; Beissbarth and Speed, 2004; Subramanian et al., 2005), which seek to identify overlaps between the identified set of genes and those from a separate, pre-defined catalog of gene sets associated with known biological functions and pathways (Cerami et al., 2011; Fabregat et al., 2018; Kanehisa and Goto, 2000; Pico et al., 2008; Wang et al., 2018a).

Paradoxically, a gene set for which there is a very strong functional enrichment may be of less interest to researchers, since the set and its function have already been well characterized by previous studies. Of greater interest are gene sets that fail functional enrichment, or overlap known functions only marginally, because it is precisely from these ‘failures’ that new biological findings emerge. In these cases, an immediate next step is to explore the biological literature, as well as complementary data sets, to learn as much as possible about the genes in question. The goal is to mine knowledge pertinent to each gene and then to use this knowledge to synthesize mechanistic hypotheses for a function that might be held in common by all genes in the set. As a result, new functional categories might be defined and added to existing collections of functions. This process of discerning relevant findings from data and literature, and reasoning on this information to synthesize functional hypotheses, has not been widely automated but is one of the central tasks performed by a genome scientist.

When reasoning about gene functions, it is critical to include knowledge of the relevant experimental conditions and biological contexts under which the gene set has been identified. For instance, FoxA family transcription factors have been found to display different roles in a very strong tissue-specific manner. FoxA regulates glucagon expression in the pancreas, GLUT2 expression in the liver, and tyrosine hydroxylase expression in dopaminergic neurons (Fox et al., 2013). Such knowledge is strikingly absent from gene set functional enrichment tools, since their computational models and databases cannot easily encode the practically infinite space of biological conditions.

Here, we sought to develop a new reasoning tool, Deep Functional Synthesis (DeepSyn), to learn the biological functions of a gene set from its totality of literature and data. Our goal was to achieve two key advances over functional enrichment: the ability to synthesize new functional hypotheses rather than rely on the predefined functions of predefined gene sets, and the ability to guide these functional hypotheses by relevant biological conditions.

Biomedical relation extraction (BRE) (Hristovski et al., 2003; Lever et al., 2019; Tsuruoka et al., 2011) can be viewed as a first attempt towards our goal. BRE is able to capture statistical associations between entities referenced in literature, such as a gene name and a disease name that are co-mentioned in many abstracts. DeepSyn extends this concept by integrating associations from literature with those from primary data, and by moving beyond individual associations to build a global knowledge graph in which to trace highly relevant pathways of associations among genes, functions, diseases and/or drugs (e.g. linking disruption of a gene to changes in the activity of proteins, pathways, and incidence of disease in relevant experimental datasets).

Defining and embedding biological functions in a global knowledge graph.

The DeepSyn knowledge graph currently captures 3,048,803 associations among 85,180 biological entities including 27,175 genes, 26,365 diseases, 4,125 drugs, and 27,515 functions (**Figure 1A**, **KEY RESOURCE TABLE**, **STAR Methods**). Gene, disease, and drug identities are defined from public databases according to standard nomenclature (Genes: HNSC; Diseases: MeSH; Drugs: Chemical name). Functions, which have been less standardized, are defined by a combination of public databases (8,542 Gene Ontology term names) and entity recognition from biomedical abstracts using a deep neural network model (yielding another 18,973 distinct biological functions). Associations among these entities are mined from biomedical abstracts (e.g. probability of co-occurrence among genes and functions or pairs of functions), ontologies (e.g. associating drugs with targets, genes with diseases) or experimental databases (e.g. associating genes via protein-protein interactions or genes and diseases through genome-wide association studies).

DeepSyn is queried with a set of genes, from which it returns the most relevant functions, diseases, and drugs extracted from the knowledge graph. The query is made context-specific by the addition of conditional constraints, which are also specified in the form of functions, diseases or drugs. Such conditions may even be specified without genes, in which case DeepSyn performs a reverse query to find the most relevant gene set (**Figure 1B**).

Results are shown in the context of an “answer graph” which is extracted from the global knowledge graph. The answer graph contains direct links from specific to general entities. Each entity in the answer is associated with a *P*-value of significance (the chance that entity would arise from random queries, **STAR Methods**), and each relationship between entities is linked to the supporting evidence in the biomedical literature or databases. For instance, a query with the disease “Inflammatory” produces an answer graph of 8 significantly associated functions and 15 genes (**Figure 1C**). In constructing the query, DeepSyn automatically adds any additional words and phrases that are frequent in the biomedical literature and contain the query or its conditional constraints. For example, the query “inflammatory” is expanded to add inflammatory myopathy, immune reconstitution inflammatory syndrome, and pelvic inflammatory disease (**Figure 1C**). Functions associated with more than one inflammatory diseases are in the answer graph, for example the top-ranked gene *PTGS2* regulates the inflammatory response by generating prostaglandins (Hata and Breyer, 2004) and plays a key role in both the treatment of pelvic inflammatory disease (Dhasmana et al., 2014) and immune reconstitution inflammatory syndrome (Shankar et al., 2007).

Functional synthesis can address multiple types of biological questions.

We formulated a panel of gold standard queries to evaluate the accuracy of associating gene sets with functions, diseases, and drugs (**Figures 1A-D**). The average performance on all of these tasks was relatively high (range 0.74 to 0.96) and represented a substantial improvement in comparison to baseline approaches based on literature co-occurrence and mutual information

(**STAR Methods**). First, we examined the ability of DeepSyn to recover the catalog of function names assigned to human gene sets by the Gene Ontology (GO) (**Figure 2A**). DeepSyn was queried with the set of genes annotated to each GO Biological Process, Molecular Function, or Cellular Component term. For each gene and condition pair, we calculated an empirical *P*-value for the corresponding knowledge graph to represent its significance compared to a random graph (**STAR Methods**). We then looked for terms among all candidate phrases ranked according to the *P*-value to genes in the term. On all three Gene Ontology categories, DeepSyn outperformed both co-occurrence and mutual information-based baseline approaches by at least 18%. Second, we evaluated the reverse function query, in which DeepSyn was given a GO term (function) name and asked to predict the corresponding set of genes. In this case, DeepSyn achieved an average accuracy of 74%, 81% and 78% for BP, CC and MF branches, again significantly outperforming baseline approaches (**Figure 2B**).

Next, we considered queries related to diseases, i.e. what disease was associated with alterations in a particular set of genes or, alternatively, what genes were associated with a particular disease. Compared to co-occurrence-based methods in the literature, we found that DeepSyn was able to outperform this baseline for at least 21% of the benchmark datasets for these two tasks (**Figure 2C**). For instance, for the disease query Fanconi Anemia (FA), 11 of the top 20 genes returned in the answer graph had been identified as FA disease genes by previous studies collected by the Monarch database (Köhler et al., 2019) (**Figure 2C**). The gene *RECQL5* was the top result, which was not documented in Monarch but was found to perform multiple functions in cells defective for the Fanconi anemia pathway (Kim et al., 2015). Further examining the answer graph, we found that *RECQL5* was connected to FA through the function of chromosome segregation (**Figure 2D**) which is well known to play a role in FA pathogenesis

(Cerabona et al., 2014). This link was present in the knowledge graph due to *RECQL5* knockdown experiments, which have shown that *RECQL5* is required for chromosome segregation through interactions with Topoisomerase IIα during mid-late S-phase (Ramamoorthy et al., 2012). Thus, DeepSyn was able to identify disease genes not covered by existing biological databases (Haendel et al., 2018). We also evaluated the reverse task, in which DeepSyn was given a set of disease genes and asked to predict the corresponding disease. On a collection of diseases in Monarch, DeepSyn achieved an average accuracy of 0.83 which also showed significant improvement compared to baseline approaches (**Figure 2E**). Finally, we evaluated the ranking of target genes for a specified drug documented in the DrugBank database (Wishart et al., 2008). We found DeepSyn was able to achieve an accuracy of 96% compared to the baseline approach of 60% (**Figure 2F**).

Providing context-specific answers to complex biological questions.

As mentioned above, gene set queries can be augmented by the specification of functions, diseases, or drugs under which the gene set has been generated or is otherwise relevant (**Figure 3A**). Such "conditional" queries can be particularly useful in the functional annotation of gene sets arising from gene or protein expression studies. As a proof of concept, we performed an mRNA expression analysis of 286 glioblastoma tumor samples from The Cancer Genome Atlas (Cancer Genome Atlas Research Network, 2008). The 40 genes with largest variance in gene expression were organized into five clusters based on the cosine similarity (**Figure 3B**). Notably, none of these clusters (each corresponding to a gene list) was deemed to be significant using standard functional enrichment against the Gene Ontology (**Supplementary Table 1**). In contrast, DeepSyn was able to synthesize significant biological functions for all gene clusters (**Figure 3B**). For example, the genes *INA, PAN-P2RY11, VSTM2A, ST8SIA3,* *PCSK2, ACTL6B and CDKN2A* formed a coherent expression cluster across glioblastoma tumor samples. Querying DeepSyn with this gene set and the condition "glioblastoma" produced an answer graph implicating a hierarchy of specific-to-general functions including vessel maturation and autophagy (**Figure 3C**).

Today, gene set enrichment remains the gold standard to measure the success of many biological discoveries in genomics. The main limitation of enrichment methods is that 1) genes perform different functions by interacting with distinct partners in different biological and clinical contexts and 2) they must rely on static manually curated gene functions. To address these limitations, here we have shown that it is possible to use machine learning algorithms to extract context-specific functional information and automatically construct accurate evaluating standards to measure the success rate of biological discoveries. This work enables a philosophical shift in functional analysis, from manual curation of literature to AI-based learning.

STAR METHODS

KEY RESOURCE TABLE

REGENT or RESOURCE	SOURCE	IDENTIFIER
Python package source code	This paper	https://github.com/wangshenguiuc/DeepSyn
Python package documentation	This paper	https://github.com/wangshenguiuc/DeepSyn
DrugBank	Wishart et al., 2008	https://www.drugbank.ca/
STITCH	Szklarczyk et al., 2016	http://stitch.embl.de/
Monarch	Haendel et al., 2018	https://monarchinitiative.org/
Human Phenotype Ontology	Schriml et al., 2019	https://hpo.jax.org/app/
MeSH	Lipscomb et al., 2000	https://www.ncbi.nlm.nih.gov/mesh
STRING	Szklarczyk et al., 2019	https://string-db.org/
inBioMap	Li et al., 2017	https://www.intomics.com/inbio/know.html
Literome	Poon et al., 2014	https://literome.azurewebsites.net/
Gene Ontology	Xiao et al., 2019	http://geneontology.org/

Collecting and processing biological databases and literature data

The core of our method is to construct a heterogeneous knowledge graph integrating both literature data and multiple public biological databases. The knowledge graph considers four types of nodes including drug, gene/protein, disease, and biological function. We collected drug-target interactions, disease-disease associations, disease-gene associations, protein-protein interactions, function-gene annotations, and associations between functions. To construct these associations, we collected the following datasets: 1) drug-target interactions from DrugBank (Wishart et al., 2008) and STITCH (Szklarczyk et al., 2016), 2) disease-gene associations from Monarch (Haendel et al., 2018), 3) disease-disease associations from Human Phenotype Ontology (Schriml et al., 2019) and MeSH (Lipscomb, 2000), 4) protein-protein interactions from STRING (Szklarczyk et al., 2019), InBioMap (Li et al., 2017) and Literome (Poon et al., 2014), and 5) function-gene annotations and function-function associations from Gene Ontology (Xiao, 2019). Besides these public databases, associations between two biological entities could also come from PubMed by using text mining techniques. We collected 16,731,155 scientific paper abstracts spanning a wide range of research areas. These abstracts were downloaded using NCBI public APIs (https://www.ncbi.nlm.nih.gov/home/develop/api/). The title of each scientific paper was appended to the corresponding abstract. The median size of each abstract was 200 words and 14 sentences. To mine this large text corpus, we first constructed a vocabulary of high quality phrases (a sequence of one or more words) which was composed from two sources: 1) drug, disease, gene, and biological function names appeared in the database network; 2) phrases mined from PubMed free text by using an unsupervised phrase mining software AutoPhrase (Shang et al., 2018; Wang et al., 2018b). The weight between two phrase nodes was calculated by their mutual information based on their co-occurring probability in the literature. Mutual information is widely adopted in the BioNLP area to handle the frequency bias for calculating the co-occurring probability (Levy and Goldberg, 2014). In addition, we only considered two phrases that co-occurred in >10 articles as literature data was noisier compared to molecular data. The weights of these edges were all normalized to the range of 0 and 1. If there were multiple edges between the same pair of nodes, the maximum value was selected as the final edge weight.

Augmenting functional associations by using deep learning

To quantify the relationship between two informative phrases, it is necessary to accurately identify whether a scientific paper is actually relevant to a particular function or not. To address this problem, we trained a neural network model based on the language corpus to explicitly predict whether a scientific paper was associated with a Gene Ontology term or not.

<u>Neural network architecture.</u> In this work, we adopted a Convolution Neural Network (CNN) with one convolution layer on top of word embeddings GloVe obtained from an unsupervised neural language model (Pennington et al., 2014). Formally, let $x_i \in R^k$ be the *k*-dimensional (*k*=50) word vector corresponding to the *i*-th word in the sentence. The filter output c_i for the *i*-th word was defined as,

$$c_i = Relu\ (\ \Sigma_{j=i}^{i+h-1} w^{(1)} x_j + b^{(1)}\ ) \tag{1}$$

Here *h* was the filter window and was set to 5 in our implementation. $w^{(1)}$ was the weight matrix of the linear transformation and $b^{(1)}$ was the bias vector. $Relu$ was the rectified linear activation function which set values below 0 to 0 (Glorot et al., 2011). We used Eq (1) as a filter to scan the sentence with length *L* and produced a feature map $c = [c_1,\ c_2,\ \ldots\ ,\ c_{L-h+1}]$. A max-overtime pooling operation then was applied over the feature map to select the maximum component c_{max} as $argmax\{c_1,\ c_2,\ \ldots\ ,\ c_{L-h+1}\}$. The pooling result was then fed into a fully connected layer with a continuous of dropout layer and softmax layer as follows,

$$o = Softmax\ (\ Dropout(\ \Sigma_{j=i}^{i+h-1} w^{(2)} c_{max} + b^{(2)}\ )\ ) \tag{2}$$

The prediction o was associated with a mean-square error loss. $w^{(2)}$ and $b^{(2)}$ were linear weights and bias of this fully connected layer, respectively.

Training algorithms and implementation details. Empirically we used the dropout rate of 0.5 and mini-batch size 500. We trained the weights of the neural network by using the gradient descent algorithm ADAM (Kingma and Ba, 2014). The learning rate of ADAM was set as 0.001. The whole computational framework was implemented using the PyTorch library (https://pytorch.org/). In total, we trained 8542 deep neural network models and got an average of AUC value 0.79 evaluated by using 5-fold cross-validation.

Training data construction. We constructed a separate learning task for each Gene Ontology term. For each predicting task, we selected all PubMed articles with the term name explicitly showing in abstracts and used these articles as positive training samples. We also selected abstracts that did not contain this term but contained sibling terms (same parent term) on the Gene Ontology as negative training samples. In this way, we encouraged neural network models to learn more discriminative representations of the language corpus by constructing a more challenging machine learning problem. For each term, we randomly select some negative samples to keep its number as ten times of the positive samples.

Constructing a biological knowledge graph.

This trained CNN classifier was then used to annotate the gene ontology term to each abstract. Such annotation significantly expands existing language models based on co-occurring probability. For instance, an article mentioned “genetic instability” in its abstract and our CNN model also predicted that the article was related to “DNA repair”. In this way, we counted ‘genetic instability’ and ‘DNA repair’ co-occur once. There were two levels of co-occurrence probability: the sentence-level and the abstract-level. The sentence-level co-occurrence probability P_{AB} between two phrases *A* and *B* was defined as,

$$P_{AB}^{sentence} = \frac{\#\ of\ sentences\ contain\ A\ and\ B}{\#\ of\ sentences\ contain\ A\ \ast\ \#\ of\ sentences\ contain\ B} \tag{3}$$

Similarly, the article-level co-occurrence probability was defined as,

$$P_{AB}^{article} = \frac{\#\ of\ articles\ contain\ A\ and\ B}{\#\ of\ articles\ contain\ A\ \ast\ \#\ of\ articles\ contain\ B} \tag{4}$$

To construct the knowledge graph, we chose to use the article-level co-occurrence probability but we also compared the performance of using sentence-level co-occurrence probability in **Figure 2**.

Querying the biological knowledge graph.

Currently, DeepSyn supported four types of user's queries including drug, disease, gene/protein, and function. Each field could be a set of entities (e.g., a gene set) or empty. If both nodes belong to the Gene Ontology or the Human Phenotype Ontology, we let the more general node point to the more specific node based on the definition of the Gene Ontology and the Human Phenotype Ontology. If two nodes are both informative phrases mined from literature, we let the node with higher occurring frequency in the literature point to the lower frequency one. DeepSyn extracted a subnetwork from the entire knowledge graph with respect to the user's query by adopting a Depth First Search (DFS) algorithm. The source nodes of the DFS algorithm were the diseases or functions in the user's query and the target nodes were genes in the user's query. We also restricted the size of the resulted subnetwork by setting the maximum layer of the network to be less than 5 and each node could be only connected to at most 10 nodes in the network. To calculate the *P*-value for each node in the returned subnetwork, we first summed up the weights of its edges in the subnetwork and then calculated an empirical *P*-value by comparing this score to a background distribution. The background distribution was fit by sampling 10,000 random subnetworks. The *P*-value of a subnetwork was thus defined as the most significant *P*-value of all nodes in this subnetwork. If one of the fields is missing, DeepSyn will automatically search for all the allowed entities in this field and return the union of all the subnetworks based on a *P*-value threshold (=0.05). For instance, if the user only searches a drug's name 'gefitinib' but does not specify any gene names, DeepSyn will automatically generate many queries, such as <"gefitinib", "*BRAC1*">, <"gefitinib", "*TP53*">..., and union all the returned subnetworks with significant *P*-values.

Evaluating the performance of DeepSyn.

To evaluate the performance of DeepSyn, we compared DeepSyn with both co-occurrence and mutual information-based baseline approaches. Both approaches were evaluated on sentence and article levels. The sentence-level and article-level mutual information between *A* and *B* were defined as,

$$MI_{AB}^{sentence} = \frac{P_{AB}^{sentence}}{P_{A}^{sentence} P_{B}^{sentence}} \tag{5}$$

$$MI_{AB}^{article} = \frac{P_{AB}^{article}}{P_{A}^{article} P_{B}^{article}} \tag{6}$$

Here $P_{AB}^{sentence}$ and $P_{AB}^{article}$ were defined by Eq. (3) and (4). $P_{A}^{sentence}$ and $P_{B}^{article}$ were the probabilities that phrase *A* showed in a sentence or an article, respectively. DeepSyn and all baseline approaches were given a biological query and conditions supported by our system. For example, the user's query could be a drug name, and all approaches were required to find associated genes. Each approach answered the query by finding the condition that had the highest co-occurrence with the query in the system. DeepSyn and all baseline approaches returned a ranking list of answers. AUROC was then used as the metric to evaluate this ranking list. When our experiments were related to GO, we evaluated the performance on all three categories of GO: Molecular Function, Cellular Component, and Biological Process. We also validated the answer to these queries according to existing biological databases.

Acknowledgments

This work is supported by NIH LM005652, TR002515, GM102365, the Chan-Zuckerberg Biohub, NIH / NIGMS P41 GM103504, and NIN / NCATS OT3 TR002026.

Competing interests

TI is co-founder of Data4Cure, Inc., is on the Scientific Advisory Board, and has an equity interest. TI is on the Scientific Advisory Board of Ideaya BioSciences, Inc., has an equity interest, and receives income for sponsored research funding. The terms of these arrangements have been reviewed and approved by the University of California San Diego in accordance with its conflict of interest policies. R.B.A. declares the following competing interests: stock or other ownership (Personalis, 23andme, Youscript); consulting or advisory role (United Health, Second Genome, Karius, UK Biobank, Swiss Personalized Health Network).

FIGURES

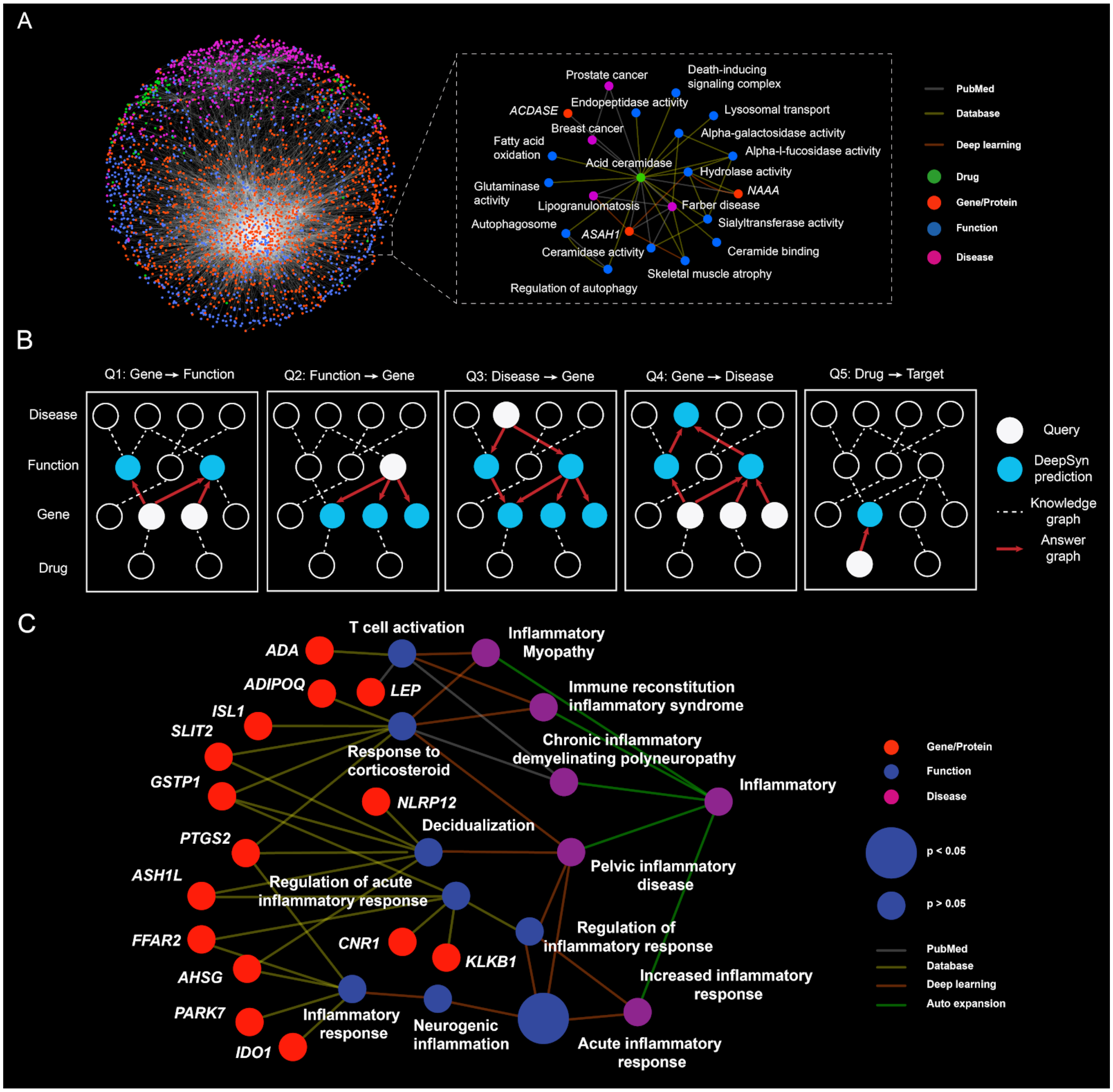

Figure 1. Overview of DeepSyn. (A) The DeepSyn knowledge network integrates a wide range of biological databases and biomedical literature. Insets illustrate the direct network neighborhoods of the drug acid ceramidase. **(B)** Five types of prediction. **(C)** The answer graph returned for the query ‘Inflammatory’.

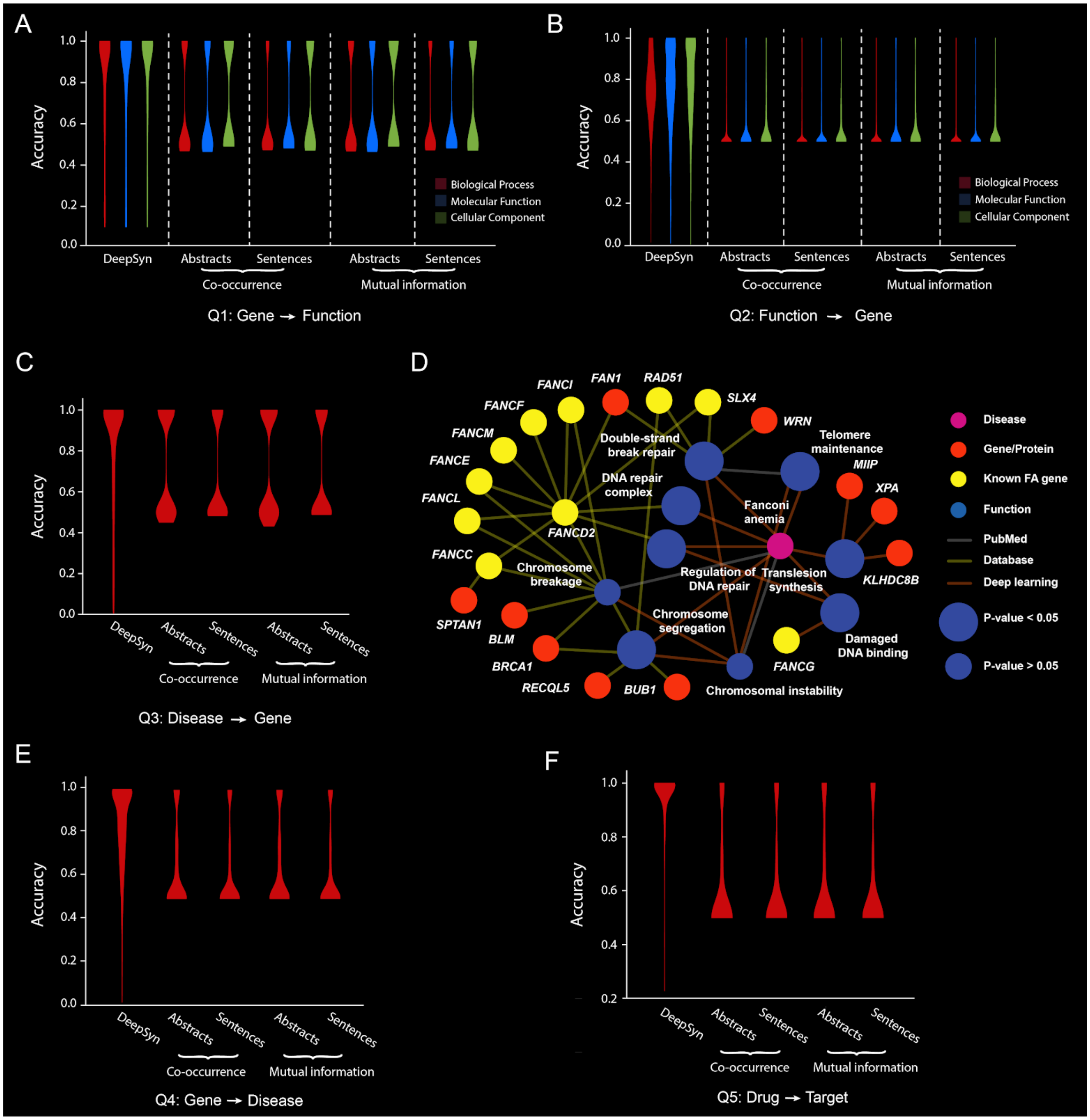

Figure 2. Prediction performance of DeepSyn. (A) Accuracy of predicting functions (Gene Ontology terms) for a given gene set. Performance is measured by Area Under Receiver Operating Characteristic (AUROC) curve. **(B)** Accuracy of predicting a gene set for a queried biological function. **(C)** Accuracy of predicting disease genes for a queried disease name. **(D)**

Answer graph and associated *P*-value returned for the query 'Fanconi Anemia'. **(E)** Accuracy of predicting a disease given a gene set. **(F)** Accuracy of predicting targets for a particular drug.

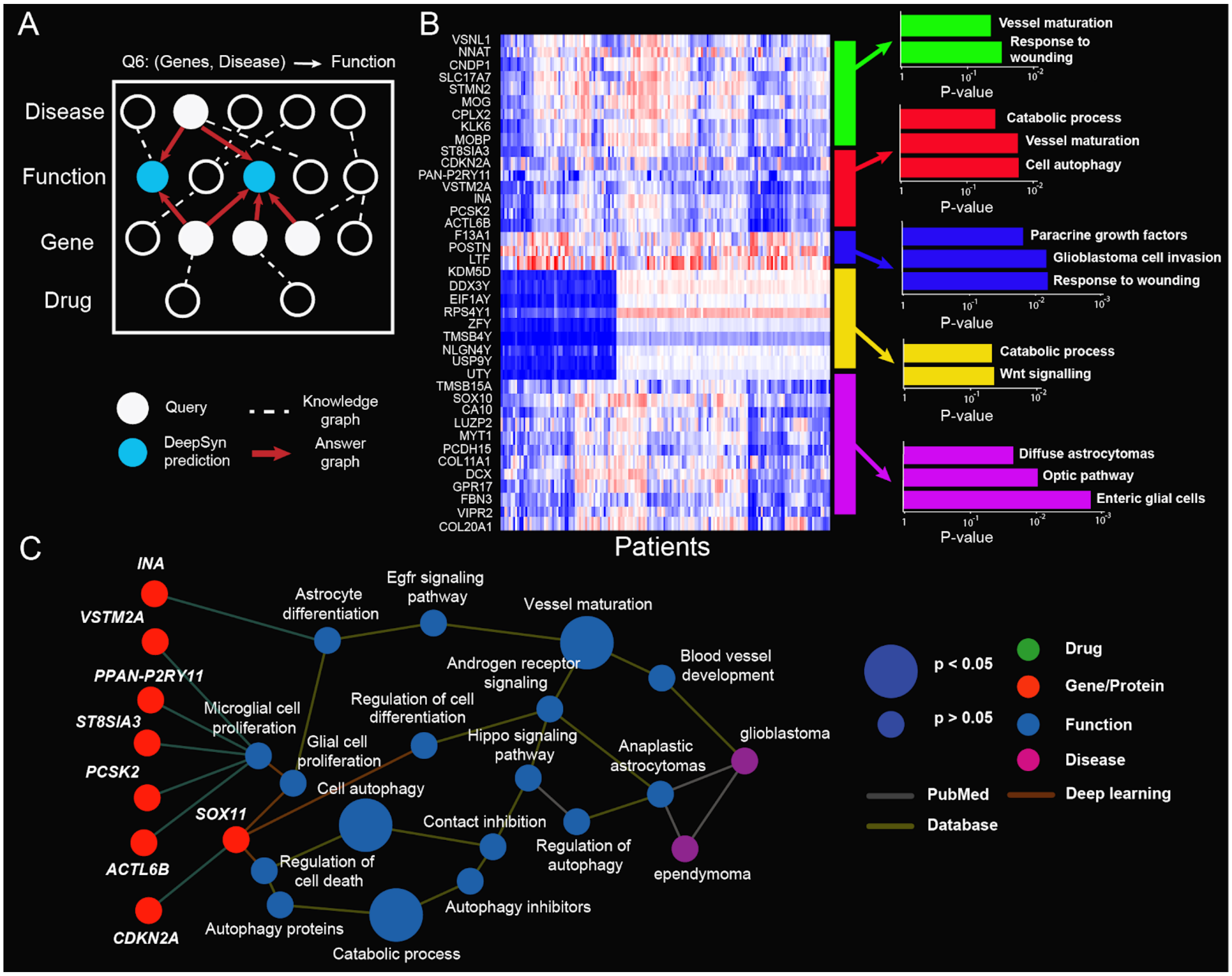

Figure 3. Annotation of genes coordinately expressed in glioblastoma. (A) DeepSyn can be used to query a gene set in the context of a disease. **(B)** gene expression clusters identified through analysis of gene expression profiles of glioblastoma tumors (The Cancer Genome Atlas Research Network 2008). Each of the gene clusters defines a gene set and is used to query DeepSyn to identify common functions. The bar plot on the right indicates the *P*-value of each

function identified by DeepSyn in the log scale. **(C)** The answer graph returned for one of the glioblastoma gene clusters consisting of seven genes (red clusters in panel **(B)**), with circle size indicating the *P*-value of significance.

REFERENCES

Al-Shahrour, F., Arbiza, L., Dopazo, H., Huerta-Cepas, J., Mínguez, P., Montaner, D., and Dopazo, J. (2007). From genes to functional classes in the study of biological systems. BMC Bioinformatics *8*, 114.

Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y.A., Müller, R., Meese, E., and Lenhof, H.-P. (2007). GeneTrail—advanced gene set enrichment analysis. Nucleic Acids Res. *35*, W186–W192.

Beissbarth, T., and Speed, T.P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics *20*, 1464–1465.

Breitling, R., Amtmann, A., and Herzyk, P. (2004). Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. BMC Bioinformatics *5*, 34.

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061–1068.

Cerabona, D., Sater, Z.A., Enzor, R., and Nalepa, G. (2014). Genomic Instability in Fanconi Anemia Results from a Combination of Chromosome Mis-Segregation in Mitosis and Unresolved Interphase DNA Damage. Blood *124*, 357–357.

Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G.D., and Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. *39*, D685–D690.

Dhasmana, D., Hathorn, E., McGrath, R., Tariq, A., and Ross, J.D. (2014). The effectiveness of nonsteroidal anti-inflammatory agents in the treatment of pelvic inflammatory disease: a systematic review. Syst. Rev. *3*, 79.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome Pathway Knowledgebase. Nucleic Acids Res. *46*, D649–D655.

Fox, R.M., Vaishnavi, A., Maruyama, R., and Andrew, D.J. (2013). Organ-specific gene expression: the bHLH protein Sage provides tissue specificity to Drosophila FoxA. Development

140, 2160–2171.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. Aistats.

Haendel, M.A., Chute, C.G., and Robinson, P.N. (2018). Classification, Ontology, and Precision Medicine. N. Engl. J. Med. *379*, 1452–1462.

Hata, A.N., and Breyer, R.M. (2004). Pharmacology and signaling of prostaglandin receptors: multiple roles in inflammation and immune modulation. Pharmacol. Ther. *103*, 147–166.

Hristovski, D., Peterlin, B., Mitchell, J.A., and Humphrey, S.M. (2003). Improving literature based discovery support by genetic knowledge integration. Stud. Health Technol. Inform. *95*, 68–73.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Kim, T.M., Son, M.Y., Dodds, S., Hu, L., Luo, G., and Hasty, P. (2015). RECQL5 and BLM exhibit divergent functions in cells defective for the Fanconi anemia pathway. Nucleic Acids Res. *43*, 893–903.

Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization.

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N.L., Matentzoglu, N., McMurry, J.A., et al. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. *47*, D1018–D1027.

Lever, J., Zhao, E.Y., Grewal, J., Jones, M.R., and Jones, S.J.M. (2019). CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nat. Methods *16*, 505–507.

Levy, O., and Goldberg, Y. (2014). Neural Word Embedding as Implicit Matrix Factorization. In Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, eds. (Curran Associates, Inc.), pp. 2177–2185.

Li, T., Wernersson, R., Hansen, R.B., Horn, H., Mercer, J., Slodkowicz, G., Workman, C.T., Rigina, O., Rapacki, K., Stærfeldt, H.H., et al. (2017). A scored human protein-protein interaction network to catalyze genomic interpretation. Nat. Methods *14*, 61–64.

Lipscomb, C.E. (2000). Medical Subject Headings (MeSH). Bull. Med. Libr. Assoc. *88*, 265–266.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

Pico, A.R., Kelder, T., van Iersel, M.P., Hanspers, K., Conklin, B.R., and Evelo, C. (2008). WikiPathways: Pathway Editing for the People. PLoS Biology *6*, e184.

Pomaznoy, M., Ha, B., and Peters, B. (2018). GOnet: a tool for interactive Gene Ontology analysis. BMC Bioinformatics *19*, 470.

Poon, H., Quirk, C., DeZiel, C., and Heckerman, D. (2014). Literome: PubMed-scale genomic knowledge base in the cloud. Bioinformatics *30*, 2840–2842.

Ramamoorthy, M., Tadokoro, T., Rybanska, I., Ghosh, A.K., Wersto, R., May, A., Kulikowicz, T., Sykora, P., Croteau, D.L., and Bohr, V.A. (2012). RECQL5 cooperates with Topoisomerase II alpha in DNA decatenation and cell cycle progression. Nucleic Acids Res. *40*, 1621–1635.

Rhee, S.Y., Wood, V., Dolinski, K., and Draghici, S. (2008). Use and misuse of the gene ontology annotations. Nat. Rev. Genet. *9*, 509–515.

Schriml, L.M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R., et al. (2019). Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res. *47*, D955–D962.

Shang, J., Liu, J., Jiang, M., Ren, X., Voss, C.R., and Han, J. (2018). Automated Phrase Mining from Massive Text Corpora. IEEE Trans. Knowl. Data Eng. *30*, 1825–1837.

Shankar, E.M., Vignesh, R., Murugavel, K.G., Balakrishnan, P., Sekar, R., Lloyd, C.A.C., Solomon, S., and Kumarasamy, N. (2007). Immune reconstitution inflammatory syndrome in association with HIV/AIDS and tuberculosis: views over hidden possibilities. AIDS Res. Ther. *4*, 29.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550.

Szklarczyk, D., Santos, A., von Mering, C., Jensen, L.J., Bork, P., and Kuhn, M. (2016). STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic Acids Res. *44*, D380–D384.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.

Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J. ’ichi, and Ananiadou, S. (2011). Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics *27*, i111–i119.

Wang, S., Ma, J., Zhang, W., Shen, J.P., Huang, J., Peng, J., and Ideker, T. (2018a). Typing tumors using pathways selected by somatic evolution. Nat. Commun. *9*, 4159.

Wang, S., Ma, J., Yu, M.K., Zheng, F., Huang, E.W., Han, J., Peng, J., and Ideker, T. (2018b). Annotating gene sets by mining large literature collections with protein networks. Pac. Symp. Biocomput. *23*, 602–613.

Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. *36*, D901–D906.

Xiao, G. (2019). Knowledge Ontology: A Tool for the Unification of Knowledge. KNOWLEDGE ORGANIZATION *46*, 236–238.

Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S., et al. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. *4*, R28.

Zhou, T., Yao, J., and Liu, Z. (2017). Gene ontology, enrichment analysis, and pathway analysis. Bioinformatics in Aquaculture: Principles and Methods 150–168.