

EnrichDO: a global weighted model for Disease Ontology enrichment analysis

Haixiu Yang^{1,*†}, Hongyu Fu^{1,†}, Meiyi Zhang¹, Yangyang Liu¹, Yongqun Oliver He², Chao Wang¹, and Liang Cheng^{1,3,*}

¹College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang 150081, China

²Unit for Laboratory Animal Medicine, University of Michigan Medical School, Ann Arbor, MI 48109, USA

³National Health Commission (NHC) Key Laboratory of Molecular Probes and Targeted Diagnosis and Therapy, Harbin Medical University, Harbin 150028, China

*Correspondence address. Liang Cheng. College of Bioinformatics Science and Technology, Harbin Medical University, 157 Nangang District, Harbin health care road, Harbin, Heilongjiang, 150081, China. E-mail: liangcheng@hrbmu.edu.cn; Haixiu Yang. College of Bioinformatics Science and Technology, Harbin Medical University, 157 Nangang District, Harbin health care road, Harbin, Heilongjiang, 150081, China. E-mail: yanghaixiu@ems.hrbmu.edu.cn

†Contributed equally.

Abstract

Background: Disease Ontology (DO) has been widely studied in biomedical research and clinical practice to describe the roles of genes. DO enrichment analysis is an effective means to discover associations between genes and diseases. Compared to hundreds of Gene Ontology (GO)-based enrichment analysis methods, however, DO-based methods are relatively scarce, and most current DO-based approaches are term-for-term and thus are unable to solve over-enrichment problems caused by the “true-path” rule.

Results: Here, we describe a novel double-weighted model, EnrichDO, which leverages the latest annotations of the human genome with DO terms and integrates DO graph topology on a global scale. Compared to classic enrichment methods (mainly for GO) and existing DO-based enrichment tools, EnrichDO performs better in both GO and DO enrichment analysis cases. It can accurately identify more specific terms, without ignoring the truly associated parent terms, as shown in the Alzheimer’s disease (AD) case (AD ranked first). Moreover, both a simulated test and a data perturbation test validate the accuracy and robustness of EnrichDO. Finally, EnrichDO is applied to other types of datasets to expand its application, including gene expression profile datasets, a host gene set of microorganisms, and hallmark gene sets. Based on the findings reported here, EnrichDO shows significant improvement via all experimental results.

Conclusions: EnrichDO provides an effective DO enrichment analysis model for gaining insight into the significance of a particular gene set in the context of disease. To increase the usability of EnrichDO, we have developed an R-based software package, which is freely available through Bioconductor (<https://bioconductor.org/packages/release/bioc/html/EnrichDO.html>) or at <https://github.com/liangcheng-hrbmu/EnrichDO>.

Graphical abstract



Keywords: Disease Ontology, human genome, enrichment analysis, double weighting

Background

Biomedical ontologies organize biomedical findings into hierarchical structures and controlled vocabularies, which are human-readable and machine-computable. Biomedical ontologies have been widely studied and applied in biomedical research and clinical practice [1]. Gene Ontology (GO) is arguably the most successful example of a biomedical ontology. It provides structured, controlled vocabularies and classifications that cover several domains of molecular and cellular biology and that are freely available for community use in the annotation of genes, gene products, and sequences [2, 3]. Disease Ontology (DO) is another representative example. DO organizes, represents, and standardizes human diseases through extensive cross-mapping and integration

of MeSH, ICD, NCI thesaurus, SNOMED CT, and OMIM disease-specific terms and identifiers [4, 5]. Other widely used biomedical ontologies, such as the Human Phenotype Ontology (HPO) [6, 7], Chemical Entities of Biological Interest (ChEBI) [8, 9], and the Ontology of Adverse Events (OAE) [10], have been developed and are included in the BioPortal [11, 12] and/or the OBO Foundry [13, 14]. With the development of biomedical ontologies, numerous excellent algorithms, tools, and platforms have emerged, including annotation, similarity calculation, enrichment analysis, and function prediction. For example, AmiGO [15] and GO-CAM [16] enable functional annotations of genes and gene products, GOSemSim provides an R package for measuring semantic similarity among GO terms and gene products [17], PANTHER is a web service for

Received: October 3, 2024. Revised: December 18, 2024. Accepted: February 14, 2025

© The Author(s) 2025. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

GO enrichment analysis [18], and PhenIX effectively diagnoses genetic diseases through computational phenotype analysis of disease-associated genomes [19]. However, most state-of-the-art methods are focused on GO, where DO-based analyses mainly focus on similarity calculation [20–22].

Biomedical ontology-based enrichment analysis can help to elucidate the potential biological significance of a particular set of genes, such as differentially expressed gene lists of high-throughput experiments. Thousands of studies have been conducted by GO-based enrichment, but DO-based enrichment methods remain relatively scarce. Osborne et al. [23] used the Unified Medical Language System (UMLS) MetaMap Transfer tool (MMTx) and the Gene Reference Into Function (GeneRIF) database to annotate the human genome with DO for the first time. LePendou et al. [24] annotated GO annotation files with DO terms using the NCBO annotator, facilitating DO enrichment analysis via a simple binomial model. The disease and gene annotations database (DGA) employs NCBO Annotator and GeneRIF to semantically annotate human genes with disease descriptors [25]. Based on the annotations of the human genome with DO, DO-based enrichment analysis is capable of discovering disease associations in high-throughput biological data. For example, KOBAS-i incorporates 5 human disease databases and provides a web server for annotation and identification of enriched diseases by binomial test and false discovery rate correction [26]. Similar web service tools include EnrichR [27], WebGestalt [28], Flame (V2.0) [29], and aGOTool [30], all of which provide disease enrichment analysis by integrating multiple disease data sources. DOSim [20] and DOSE [31] are R packages widely used for enrichment analysis that employ the hypergeometric model and gene set enrichment analysis. The web tool ADEPTUS, which is based on high-quality curated databases with information on gene expression profiles and diseases, enables various functional genomics analyses, including DO enrichment analysis [32].

Obviously, DO-based annotation and enrichment are relatively scarce, and current DO enrichment methods are lacking. On one hand, gene–DO annotations and DO terms of the current methods are outdated, and some web services for DO enrichment are inaccessible due to lack of maintenance. On the other hand, these methods are based on classic models, such as the hypergeometric test, Fisher exact test, the χ^2 test, and the binomial test, which do not take dependencies between DO terms into consideration. DO has a hierarchical structure that forms a directed acyclic graph (DAG) that follows the “true-path” rule, which means if a gene is annotated to node *t*, it is also annotated to all parent terms of *t*. This inheritance problem can lead to some terms being over-enriched, which has been well addressed in GO-based enrichment methods [33, 34].

In this study, we annotated the human genome with DO, utilizing the latest data source, and developed a novel DO enrichment method, EnrichDO. EnrichDO tackled the “inheritance problem” moderately by considering the DO graph topology on a global scale and double-weighted the annotated genes. We compared EnrichDO with classic enrichment methods and current DO-based enrichment tools and evaluated the performance and robustness of EnrichDO. We also performed EnrichDO on datasets directly related to disease (including a microarray gene expression dataset and an RNA sequencing expression dataset), as well as the datasets indirectly related to disease (including a microorganism coexpressed host gene set and biological processes-related datasets). Our results showed that EnrichDO outperformed classic methods and current DO-based enrichment tools, demonstrating the effectiveness of EnrichDO for gaining insight into the signif-

icance of a particular set of genes in the context of disease. EnrichDO has been implemented as an R-based tool, which is available in the Bioconductor EnrichDO project page [35] and GitHub [36].

Materials and Methods

Data collection and preprocessing

Standard DO terms were downloaded from the DO Database [37] (data-version: releases/2024-03-28/doid.obo). The DO database is a comprehensive resource that aims to organize, represent, and standardize human diseases. The DO semantically integrates disease and medical vocabularies through extensive cross-mapping of DO terms to MeSH, ICD, NCI thesaurus, SNOMED CT, and OMIM disease-specific terms and identifiers [4, 5]. The DO has been widely used for disease annotation by various biomedical databases [5]. The obo file records the information of 14,019 DO terms, and we collected information on 11,537 DO terms, excluding obsolete DO terms (*is_obsolete*=true). The names and synonyms of DO terms were extracted to construct the dictionary of disease names for DO annotation. The DOID and *is_a* relations were used to construct a DAG of DO.

Information on human genes with disease descriptions was downloaded from GeneRIF (version date: 31 March 2024). GeneRIF is a database that provides concise descriptions of gene function, making it the most suitable resource for DO annotation to infer gene–disease associations [23]. GeneRIF is available from the NCBI Gene Database [38]. GeneRIF entries were manually extracted from scientific literature, including Tax IDs, Gene IDs, PubMed IDs, and concise GeneRIF textual descriptions (up to 250 characters) of gene function. GeneRIF entries of *Homo sapiens* were extracted from the *generifs_basic* file, and GeneRIF text was annotated with DO terms.

Annotating the human genome with DO terms

Semantically annotating the human genome with DO terms can connect biomedical gene and disease data through the lens of human disease. In this study, the textual descriptions of gene function in the GeneRIF database were annotated with DO terms using the University of Michigan’s Mgrep tool. Mgrep is an efficient tool for mapping free text to ontology terms [39], which is used in the concept recognition step by the NCBO Annotator [11]. We choose Mgrep because it is claimed to be a fast and scalable tool for concept recognition and has a high degree of customizability vis-à-vis dictionaries and resources [40]. For concept recognition, the data resources represent a particular type of biomedical knowledge, and a dictionary represents a set of terms to recognize in the biomedical data resource. In our study, the biomedical data resources are GeneRIF entries of *H. sapiens*; the disease dictionary was constructed by extracting all standard DO terms and their synonyms. The concept recognizer identifies the related disease name in GeneRIF entries and maps it to the concept of DO terms in the dictionary. Last, 525,289 out of 1,039,774 GeneRIF entries of *H. sapiens* were annotated with 4,407 disease terms.

Data preprocessing and result filtering were conducted as follows: (i) the disease dictionary was generated to include all standard DO terms and synonyms, with character length no less than 3, and to exclude obsolete DO terms (*is_obsolete*=true); (ii) only GeneRIF entries of protein-coding genes (pc genes) of *H. sapiens* were retained; (iii) disease names annotated to DO terms and synonyms were standardized to DOIDs; and (iv) all repeating results were removed. As a result, we acquired 191,542 gene–DO annota-

Table 1: Annotations of molecules with ontology terms in EnrichDO

Level	Number of DO	Number of annotated genes
1	1	15,106
2	8	14,400
3	120	15,185
4	212	15,526
5	574	15,929
6	1,242	15,790
7	1,009	13,686
8	850	11,600
9	535	7,826
10	170	3,231
11	58	1,400
12	26	567
13	8	186

tions, which involved 4,361 DO terms and 15,106 pc genes. The DO annotations describe unique roles for human genes in the context of disease and can be used for DO enrichment analysis.

Semantic expansion in DO DAG

We leveraged the hierarchical structure of DO to expand annotations. Like GO, DO has a hierarchical structure that forms a DAG, which follows the “true-path” rule. The additional annotation information is produced using the semantic relationships in DO, such as “is_a” relation, as follows: first, DOID, name, and “is_a” relations were used to construct the DAG. Every DO term is represented by a node in the graph, while the “is_a” relation is represented by an edge in the graph. Two terms with an “is_a” relation are represented as a child node and a parent node in the graph; a child can have multiple parents. For example, “B-cell acute lymphoblastic leukemia” (DOID:0080638) is_a “acute lymphoblastic leukemia” (DOID:9952) and “lymphoma” (DOID:0060058); the node “B-cell acute lymphoblastic leukemia” is a child of “acute lymphoblastic leukemia” and “lymphoma” and is a more specific biological classification than its parents. The term “disease” (DOID:4) is the root node of the DAG, which has no parents, and is the most general node. Next, according to the gene–DO annotations mentioned above, a set of genes annotated with each DO term in the DAG were also annotated to its parents and its ancestors (the so-called true-path rule). Leaf nodes with no annotated genes were iteratively pruned, until all nodes in the final DAG tree contained at least 1 annotated gene. After pruning, the DAG contained 4,813 nodes. Last, all nodes of DAG were marked as different levels. For a node n , the level was defined as the length of the longest path from the root to node n . The root node, the term “disease,” was marked as level 1; level 13 is the highest level. Nodes at the same level do not share any edges and can be investigated independently. The levels of DAG and the statistics of DO terms are displayed in Table 1.

DO enrichment based on weighted algorithm

Classical enrichment analysis

Overrepresentation analysis (ORA) is a widely used term-for-term approach for enrichment analysis. The ORA method measures the statistical proportion of a preselected list of genes of interest (e.g., differentially expressed gene list) and specific gene sets according to known gene functions (e.g., DO terms). The hypergeometric test

is a classical ORA test, wherein the P value indicates the probability of the null hypothesis, which can be calculated as follows:

$$p = 1 - \sum_{k=0}^{r-1} \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}} \quad (1)$$

The value N is the number of all human genes annotated to DO terms, while m indicates the number of genes annotated to term t , and n indicates the size of the interesting gene list, of which r are included in the term t . A small P value indicates a low probability of randomly obtaining such statistical proportion between the interesting gene list and term t . A multiple testing correction method, such as the Benjamini–Hochberg method, was then used to adjust the P value to control the type I error (false-positive) rate.

Weighted DO enrichment analysis

The disadvantage of the classical enrichment analysis approach is that it ignores the annotation dependencies between DO terms that are caused by the “true-path” rule. This inheritance problem can lead to some terms being over-enriched. To address this problem, we developed the weighted DO enrichment analysis method, referred to as EnrichDO, by double-weighting the annotated genes and integrating DO graph topology on a global scale. On the one hand, we assigned distinct initial weights to directly annotated genes and to indirectly annotated genes caused by the “true-path” rule, respectively, with the intention of reinforcing the saliency of direct gene–DO annotations while reducing the influence of indirectly annotated genes. On the other hand, we dynamically down-weighted genes in less significant nodes to reinforce differences in significance between the parent and its children, as described by Alexa et al. [33].

We first define 4 key concepts:

(i) *Initial weight* w_i . Initially, for each DO term, the weights for all directly annotated genes are set to 1, and the weights for indirectly annotated genes decrease by 0.1 for each level inherited upward. The smaller the weights of indirectly annotated genes, the less the contribution that these genes have to the enrichment analysis. Notably, smaller weights can cause certain nodes to be missed during the enrichment analysis. In this study, the weights for indirectly annotated genes were no less than 0.5. Decreasing the initial weights of the indirectly annotated genes, which are typically annotated with parent or ancestor nodes, can effectively alleviate the over-enrichment problem.

(ii) *Dynamic weight* w_d . For a given DO term t and its child c , the weight assigned to genes annotated to this pair of terms is defined as follows:

$$w_d = \frac{\log(\text{score}(c))}{\log(\text{score}(t))} \quad (2)$$

where $\text{score}()$ is the P value of the hypergeometric test with weighted genes. The weights for genes annotated to each node are memorized and updated during the process. Dynamic weight reinforces the differences in significance between the parent and its children, thereby reducing the local impact of inheritance.

(iii) *Penalty score*. The penalty score is a penalty value for gene weights used to supplement the initial weight and dynamic weight. It is defined as follows:

$$\text{penal} = \max\left(\frac{1}{10} \times \frac{\log(\text{xmin})}{\log(\text{score}(t)) + \log(\text{score}(c))}, 1\right) \quad (3)$$

where x_{min} is the minimum positive number in R , $1/10$ is the scale factor, and $score()$ is the P value of the hypergeometric test with weighted genes mentioned above. When the P values of the parent and child nodes are similar, the penalty score significantly decreases the gene weights, rather than w_d .

(iv) **Significance score.** The significance score for a given DO term t is calculated by applying the hypergeometric test with weighted genes. The number r in formula (1), which indicates the number of interesting genes included in term t , is replaced by rounding down the sum of the weight of r genes. The number m , which indicates the number of genes annotated to term t , is replaced by rounding down the sum of the weight of m genes.

EnrichDO is a double-weighted iterative model. Given a threshold of 0.01 (or less), our aim is to find significantly enriched DO terms with P values < 0.01 . The overall process of the algorithm is shown in Fig. 1A, and the detailed process is described step-by-step as follows:

Step 1. Given the DAG with levels (mentioned in “Semantic expansion in DO DAG”), we set distinct initial weights for annotated genes for each node. We process the nodes bottom-up, from the highest level, and then iteratively move to nodes at lower levels, with the aim of identifying the most specific nodes with minimum required significance (as shown in Fig. 1A, B). For the current node t , all its children have significance scores; children with P values larger than the threshold are excluded, as shown in Fig. 1A and B. The node t is then processed by step 2 to step 5, as shown in Fig. 1C.

Step 2. For the current node t , we calculate its significance score, using concept (iv) described above. If the score is larger than the threshold, then the process moves to the next node. Otherwise, t is compared to each of its children to discover the most significant nodes locally. In detail, for each child c , we calculate the weight w_d for the pair of terms. When $w_d > 1$, it indicates that c is more significant than t ; thus, node c is a local optimum, and c is moved from children to sigChildren. Otherwise, c remains in children. This step is shown in Fig. 1C-(1).

Step 3. For each child c in sigChildren, the weight of the genes annotated to c is decreased, by dividing term t and all its ancestors by w_d and the penalty score. The child c is then removed from sigChildren, until the sigChildren is empty. This step is shown in Fig. 1C-(2).

Step 4. Steps 2–3 are executed recursively. The significant score of t is recalculated with updated weights and is compared to its remaining children. This process is carried out until either of the 2 following situations arises: (i) there are no remaining children, in which case the process moves to the next node, or (ii) remaining children exist, in which case the process moves to step 5.

Step 5. For each remaining child c , it is less significant than t . Thus, node t is a local optimum, and the weight of genes annotated to c is decreased by multiplying w_d ($w_d < 1$) and dividing it by the penalty value, thereby reducing the contribution of genes to the child nodes. The significance score is then recalculated, and the process moves to the next node. This step is shown in Fig. 1C-(3).

Last, the multiple testing correction procedure from the Benjamini–Hochberg method (1995) is applied to adjust the significance score (P values).

Results

Summary of annotations of the human genome with DO

Annotating human genes with DO is crucial to advancing the discover of gene–disease associations. In this study, GeneRIF and DO terms were utilized to annotate the human genome with DO, using the Mrep tool. We annotated 525,289 GeneRIF entries of *H. sapiens* with 4,407 disease ontologies directly. We then integrated and filtered the annotated results, finally acquiring 191,542 gene–DO annotations with 4,361 DO terms and 15,106 pc genes. We analyzed all gene–DO annotations except for the term “disease” (DOID:4), which lacks specific significance. The number of genes annotated with each DO term was distributed from 1 to 4,730. The annotation results show an uneven distribution (Fig. 2A). Out of 4,360 DO terms, 2,971 (68.14%) have fewer than 10 annotated genes, among which 1,217 specific terms have only 1 annotated gene. In contrast, 29 DO terms have more than 1,000 annotated genes. The DO term “cancer” (DOID:162), with the highest number of annotated genes (4,730), is a disease of uncontrolled cellular proliferation that is malignant and primary, characterized by both local cell invasion and metastasis. It is a generalized and widely studied disease located at level 4 of DAG. Specific cancer types, such as breast cancer, hepatocellular carcinoma, and colorectal cancer, are also widely studied and annotated with more than 3,000 genes.

The number of DO terms annotated by each gene was distributed from 1 to 506, which presented an uneven distribution, as shown in Fig. 2B. Of 13,699 genes, 213 (1.55%) were annotated to more than 100 DO terms, while 9,595 genes were annotated to no more than 10 DO terms, of which 2,463 genes annotated to only 1 DO term. The pc gene “tumor necrosis factor (TNF),” which annotated with the highest number of DO terms (506), encodes a multifunctional proinflammatory cytokine that belongs to the TNF superfamily. This cytokine is involved in the regulation of a wide spectrum of biological processes, including cell proliferation, differentiation, apoptosis, lipid metabolism, and coagulation. It has been implicated in a variety of diseases, including autoimmune diseases, insulin resistance, psoriasis, rheumatoid arthritis, ankylosing spondylitis, tuberculosis, autosomal dominant polycystic kidney disease, and cancer. Mutations in TNF affect susceptibility to cerebral malaria, septic shock, and Alzheimer’s disease [41]. Other notable genes, such as interleukin 6 (IL6), tumor protein p53 (TP53), vascular endothelial growth factor A (VEGFA), transforming growth factor beta 1 (TGFB1), and matrix metalloproteinase 9 (MMP9), are also implicated in a variety of diseases and annotated to more than 400 DO terms.

High annotation quality is essential to the performance of enrichment analysis. We manually validated the annotations and compared them with the disease and gene annotations of the DGA database [25]. We randomly extracted and manually checked 5,000 annotation mappings, excluding the term “disease” (DOID:4). Out of 5,000 mappings, 204 were incorrectly annotated, and the false positive (4.08%) was within an acceptable range, which confirmed the accuracy of the annotations of EnrichDO. In addition, we compared the annotations with the DGA database, which was widely used for DO enrichment analysis [31]. We selected 4 diseases with annotated gene numbers in different distribution intervals—namely, Leigh disease (DOID:3652, with 53 annotated genes), retinitis pigmentosa (DOID:10584, with 158 annotated genes), colorectal carcinoma (DOID:0080199, with 549 annotated genes), and schizophrenia (DOID:5419, with 1,299 annotated

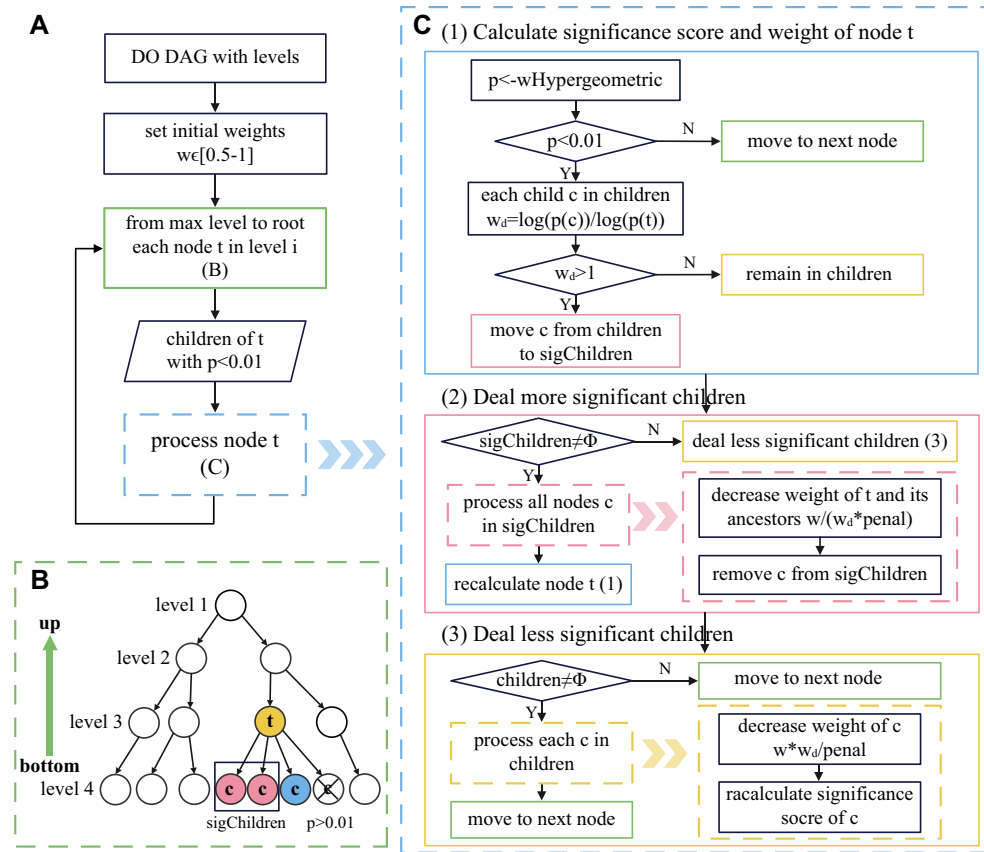


Figure 1: Flowchart of EnrichDO. (A) The main process of EnrichDO. (B) The nodes of the DAG were processed bottom-up, and the children of node **t** were divided into different sets according to their significance scores. (C) Each node **t** was processed in detail step by step.

genes). We then collected the manually curated disease–gene relations for these 4 diseases from the DisGeNET [42], MalaCards [43], and DISEASE [44] databases as standards. When comparing the directly annotated genes of Leigh disease against these 3 databases, we acquired 27, 11, and 26 overlaps, respectively; overlaps with DGA were 14, 7, and 13 (see Fig. 2C). Similar results were observed for retinitis pigmentosa, colorectal carcinoma, and schizophrenia. As shown in Fig. 2D–F and Table 2, EnrichDO annotations consistently outperformed those of the DGA database in overlap counts, demonstrating the accuracy of EnrichDO annotations and their reliability for DO enrichment analysis.

Algorithm comparison and assessment

Comparison with other methods

We compared EnrichDO with the classic ORA method (common hypergeometric test) and topGO [33], both of which have been widely used in GO enrichment analysis. Consequently, we conducted comparisons based on GO and DO separately, utilizing the latest annotations of GO and DO. For the topGO analysis, the parameters for the algorithm and statistic were set to “weight” and “fisher,” respectively.

We first applied 3 methods to perform GO (biological process, BP) enrichment analysis on acute lymphoblastic leukemia (ALL) [45], which had been previously analyzed using topGO. The dataset consists of 95 patients with B-cell ALL and 33 patients with T-cell ALL in the study. A total of 194 differentially expressed pc genes with $|\log FC| \geq 1$ were obtained for the enrichment analysis. The top 10 enriched GO terms of EnrichDO are displayed in Fig. 3A. Among the top 10 results, EnrichDO had 4 terms that

overlapped with topGO, including “T-cell receptor signaling pathway” (GO:0050852), “adaptive immune response” (GO:0002250), “peptide antigen assembly with MHC class II protein complex” (GO:0002503), and “B-cell activation” (GO:0042113). The top enriched GO term identified by EnrichDO was “antigen receptor-mediated signaling pathway” (GO:0050851), which was defined as the series of molecular signals initiated by the cross-linking of an antigen receptor on a B or T cell. The hypergeometric test had only 1 overlap with EnrichDO, named “immune response-regulating cell surface receptor signaling pathway” (GO:0002768). Additionally, GO terms such as “positive regulation of T-cell activation” (GO:0050870) and “immune response-activating cell surface receptor signaling pathway” (GO:0002429) were also top-ranked in EnrichDO’s results. As shown in Supplementary Table S1, the enriched terms of EnrichDO and topGO are more specific than those of the hypergeometric test. Notably, although the results of topGO are specific, they overlook the parent nodes of some significant nodes, which was more evident in the following DO enrichment analysis.

We then performed DO enrichment analysis on a publicly available dataset for Alzheimer’s disease (AD), which is a neurological disorder. AD is associated with declines in thinking, memory, and language and with personality changes and changes in the brain that eventually result in the loss of ability to carry out simple daily tasks [46]. Due to the lack of a gold-standard dataset for DO enrichment analysis, we collected AD-related genes from DisGeNET [42], which is deemed a gold-standard dataset. We collected 91 curated disease-associated pc genes from the DisGeNET, which were taken as interesting gene set for DO enrichment analysis.

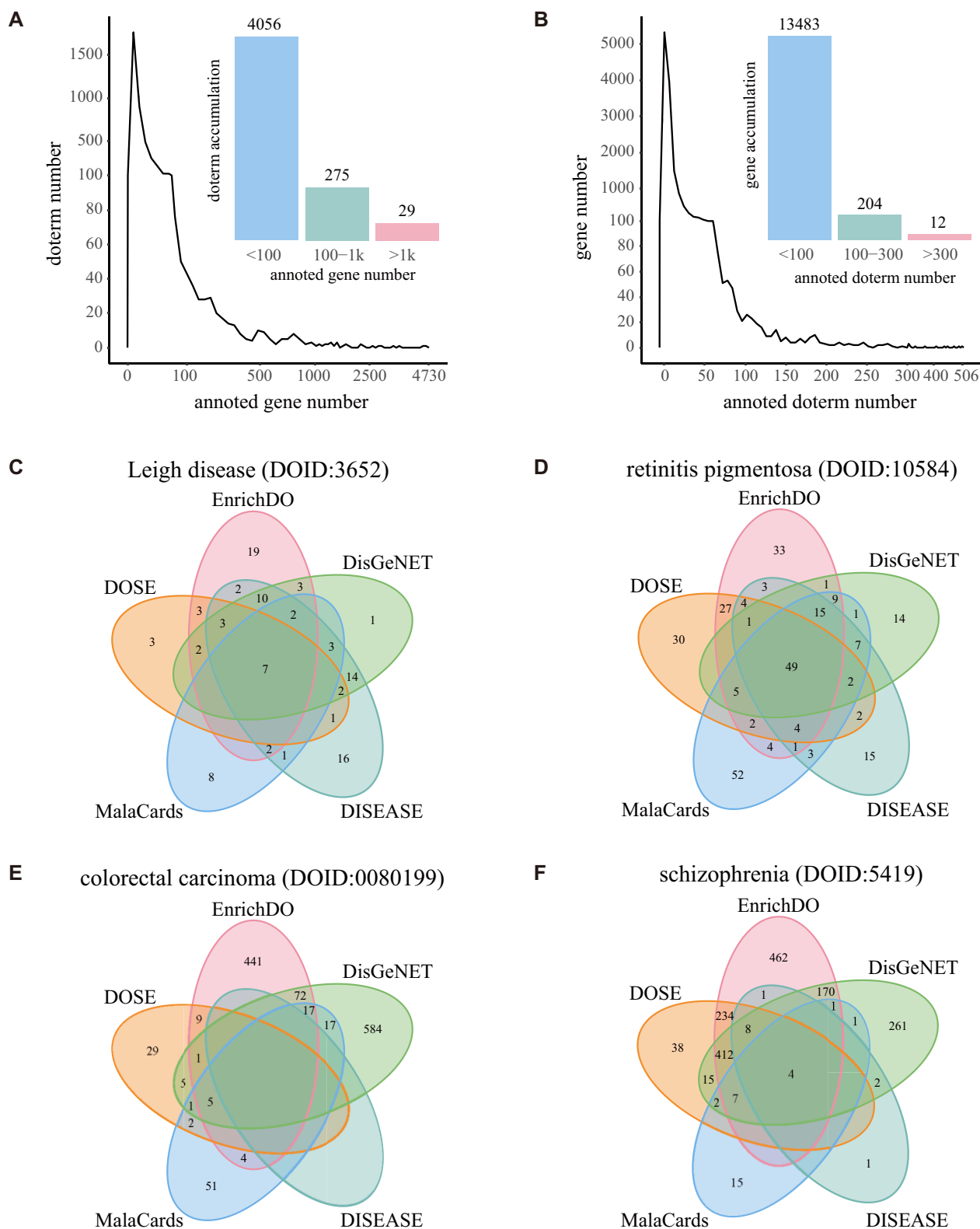
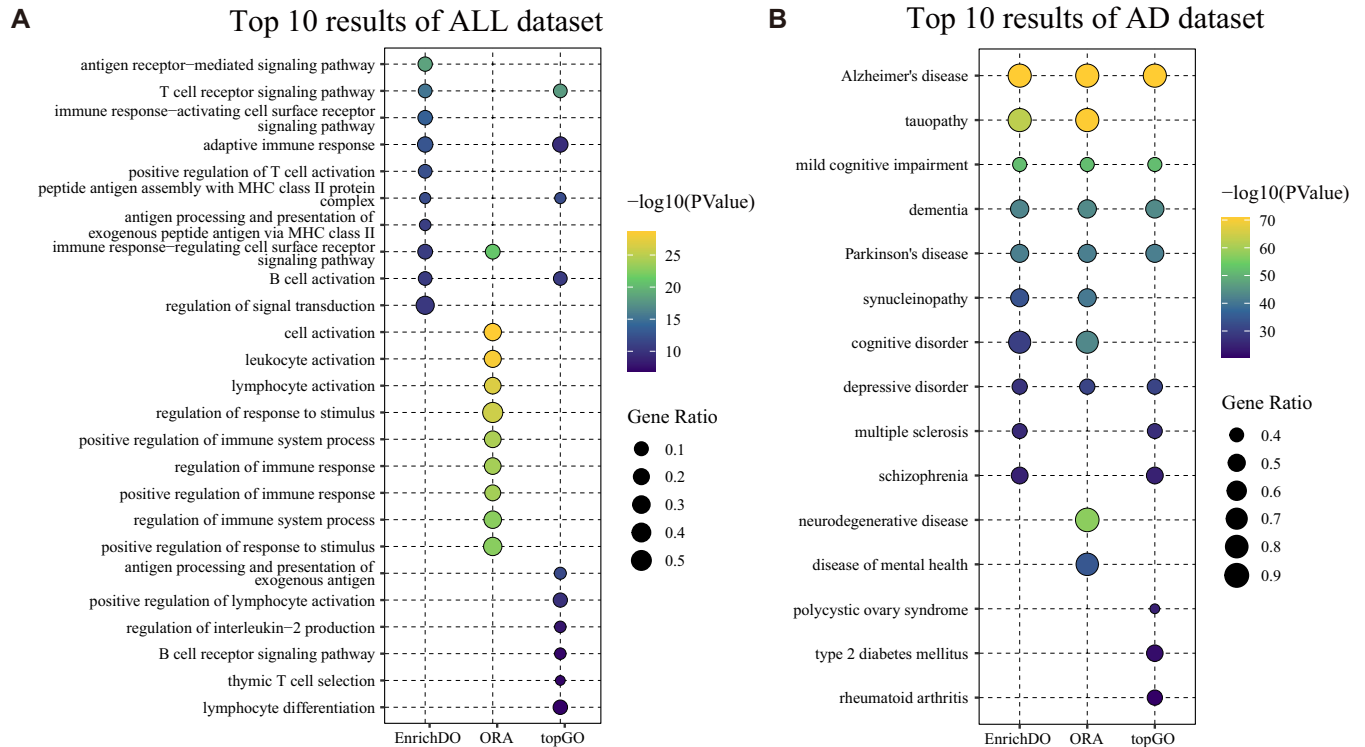


Figure 2: Statistics and comparisons of annotations of the human genome with DO. (A) Statistics of genes annotated with each DO term. (B) Statistics of DO terms annotated to each gene. (C) Annotations of Leigh disease in different databases. (D) Annotations of retinitis pigmentosa in different databases. (E) Annotations of colorectal carcinoma in different databases. (F) Annotations of schizophrenia in different databases.

Table 2: Comparison of annotations between EnrichDO and DGA for different DO terms

DOID	Disease	Number of related genes in EnrichDO and DGA*	Database	Number of related genes	Number of overlaps with EnrichDO	Number of overlaps with DGA
DOID:3652	Leigh disease	53/21/15	MalaCards	23	11	7
			DISEASE	63	26	13
			DisGeNET	47	27	14
DOID:10548	Retinitis pigmentosa	158/126/92	MalaCards	154	89	62
			DISEASE	106	77	62
			DisGeNET	104	80	57
DOID:0080199	Colorectal carcinoma	549/52/15	MalaCards	97	26	8
			DISEASE	0	0	0
			DisGeNET	702	95	12
DOID:5419	Schizophrenia	1299/720/665	MalaCards	30	12	13
			DISEASE	16	13	12
			DisGeNET	883	602	448

*The three values in the third column represent the number of genes annotated with the corresponding DO term in EnrichDO and DGA, as well as the number of overlaps between EnrichDO and DGA.

**Figure 3:** Comparison of EnrichDO with other methods. (A) Comparison of GO enrichment analysis of EnrichDO, topGO, and the hypergeometric test on the ALL case. (B) Comparison of DO enrichment analysis of EnrichDO, topGO, and the hypergeometric test on the AD case.

All 3 methods were based on the latest annotations of DO from EnrichDO.

The top 10 enriched DO terms of EnrichDO are presented in Fig. 3B, with ascending order of their computed P values. The first-ranked disease is AD, indicating that EnrichDO can accurately enrich the specific disease. The second-ranked DO term is “tauopathy” (DOID:680), a heterogeneous group of neurodegenerative diseases characterized by aggregated tau proteins, which is the parent of AD [47]. “Mild cognitive impairment” (DOID:0080832) and “dementia” (DOID:1307) were also top-ranked. Mild cognitive impairment (MCI) is an intermediate state between normal cogni-

tion and dementia. MCI is a useful label in clinical settings to help identify individuals who are at risk of developing AD [48]. Dementia is characterized by progressive deterioration in cognition, function, and behavior; AD is the most frequent cause of dementia [48]. Thus, top 10 ranked DO terms are all neurological or cognitive system diseases, indicating the accuracy of the enrichment algorithm.

Results from the comparison with the hypergeometric test and topGO on AD are displayed in Fig. 3B and Table 3. All 3 methods acquired low P values due to the manually curated AD-related interesting genes. In the top 10 results, topGO had 7 overlaps with

Table 3: Comparison of enrichment results with other methods on the AD case

DOID	DOTerm	Level	EnrichDO		ORA*		topGO	
			P	Rank	P	Rank	P	Rank
DOID:10652	Alzheimer's disease	7	2.19E-71	1	2.44E-71	1	2.44E-71	1
DOID:680	Tauopathy	6	4.38E-63	2	3.89E-71	2	1	4054
DOID:0080832	Mild cognitive impairment	4	3.09E-52	3	3.09E-52	4	3.09E-52	2
DOID:1307	Dementia	4	3.02E-43	4	4.52E-44	5	4.52E-44	3
DOID:14330	Parkinson's disease	7	3.27E-42	5	3.48E-42	7	3.48E-42	4
DOID:0050890	Synucleinopathy	6	2.26E-34	6	3.30E-41	8	1	3178
DOID:1561	Cognitive disorder	3	1.26E-30	7	1.01E-43	6	5.72E-14	45
DOID:1596	Depressive disorder	5	2.19E-28	8	1.40E-31	10	1.40E-31	5
DOID:2377	Multiple sclerosis	8	8.99E-27	9	1.33E-27	12	1.33E-27	6
DOID:5419	Schizophrenia	5	2.70E-25	10	2.78E-25	16	2.78E-25	7
DOID:1289	Neurodegenerative disease	5	1.68E-22	15	8.17E-58	3	1	4348
DOID:150	Disease of mental health	2	2.14E-11	107	3.61E-35	9	1	4810
DOID:1596	Depressive disorder	5	2.19E-28	8	1.40E-31	10	1.40E-31	5
DOID:11612	Polycystic ovary syndrome	7	3.05E-25	11	3.05E-25	17	3.05E-25	8
DOID:9352	Type 2 diabetes mellitus	7	2.02E-22	16	2.07E-22	28	2.07E-22	9
DOID:7148	Rheumatoid arthritis	8	2.96E-21	18	3.31E-21	33	3.31E-21	10

*ORA represents the hypergeometric test.

EnrichDO's results; the DO terms "tauopathy" (second) and "synucleinopathy" (sixth) were identified as nonsignificant (P value = 1) by topGO, while "cognitive disorder" (seventh) ranked 45th in the topGO results. The more specific DO terms "polycystic ovary syndrome," "type 2 diabetes mellitus," and "rheumatoid arthritis" were also among the top 10 results but are not as closely related to AD as are the aforementioned 3 terms. The hypergeometric test had 8 DO terms that overlapped with EnrichDO, and the top 2 enriched DO terms were "Alzheimer's disease" and "tauopathy." Nevertheless, DO terms ranked differently between the 2 methods, with the top-ranked DO terms of EnrichDO being more specific. The phenomenon is more evident in other comparisons of the significant results. Setting the threshold $P < 0.01$ and $P\text{-adjust} < 0.01$, the hypergeometric test identified 859 significant terms, far exceeding those identified by EnrichDO (405) and topGO (272). We investigated the top 50, 100, and 200 results of the 3 methods in the DAG. The results of the hypergeometric test are more general, are located at lower levels, and primarily converge on same branches. Conversely, the results of topGO are more specific, located at higher levels, and distributed across distinct branches (Supplementary Table S2). The above results indicate that the hypergeometric test performed enrichment term-for-term and cannot solve the over-enrichment problem, while topGO recognized more specific terms but might overlook truly relevant nodes. In contrast, EnrichDO often yields more accurate, and more specific DO terms, thereby overcoming the over-enrichment problem. At the same time, it is also a moderate weighted algorithm that will not neglect true significant terms, even though their child nodes are significant.

Comparison with current DO-based enrichment analysis tools

We also applied the AD case to compare EnrichDO with DOSE (v3.26.2), Flame (V2.0), and KOBAS-i, which are widely used disease-based enrichment analysis tools. The enrichment tool option for Flame was set to "aGoTool," and the minGSSize and maxGSSize for DOSE were set to 5 and 5,000, respectively. All other parameters for these tools were set to default. Among the top 10 results, DOSE had 4 overlaps with EnrichDO. Notably, it recognized "tauopathy" as the most significantly enriched DO term, which

is the parent of "Alzheimer's disease." The latter term ranked second. The parent and ancestors of the term "tauopathy," including "neurodegenerative disease," "central nervous system disease," and "nervous system disease," were also significantly enriched (Fig. 4). This may be attributed to the inheritance problem caused by the "true-path" rule. DOSE identified statistically significant enriched diseases based on the hypergeometric test, which did not take topological relationships between DO terms into consideration. Another explanation for the results is that the annotation data of DOSE has not been updated and therefore contains only 10,312 annotated genes, compared to 15,106 pc genes in EnrichDO. Furthermore, DOSE obtains similar results when the enrichment background is restricted to pc genes (Supplementary Table S3). Flame had 3 overlaps with EnrichDO: "AD," "dementia," and "cognitive disorder." Flame recognized AD as the top correlated disease; however, in comparison to EnrichDO, other results obtained by Flame were more general, as shown in Fig. 4. In addition, the branches of Flame results are more similar to DOSE (with 6 overlaps), and there is a branch related to metabolic diseases. KOBAS recognized 5 duplicate AD-like diseases, with different disease names that come from different human disease databases. The other results were mainly related to cardiovascular disease and metabolic disease. KOBAS incorporated 3 human disease databases, OMIM, KEGG DISEASE, and the NHGRI GWAS Catalog (NHGRI) but did not include DO. Multiple data sources might lead to redundancy in the enrichment results. The above results indicated that EnrichDO performed better than current methods.

Robustness of EnrichDO

We tested the robustness of EnrichDO with the AD dataset by removing interesting genes and by adding noise, respectively. First, we removed the AD dataset at 5% intervals and repeated the EnrichDO method 100 times for each removal. The number of significant overlapped DO terms (top 100) decreased slowly compared to the original data, and the ratio of overlapped DO terms to original significant DO terms remained at 74.17%, even after removal of up to 30% of the AD dataset (Fig. 5A). These results indicate that the EnrichDO method is robust to data removal. In a similar way, noise was added to the AD dataset. We added 5% to 30%

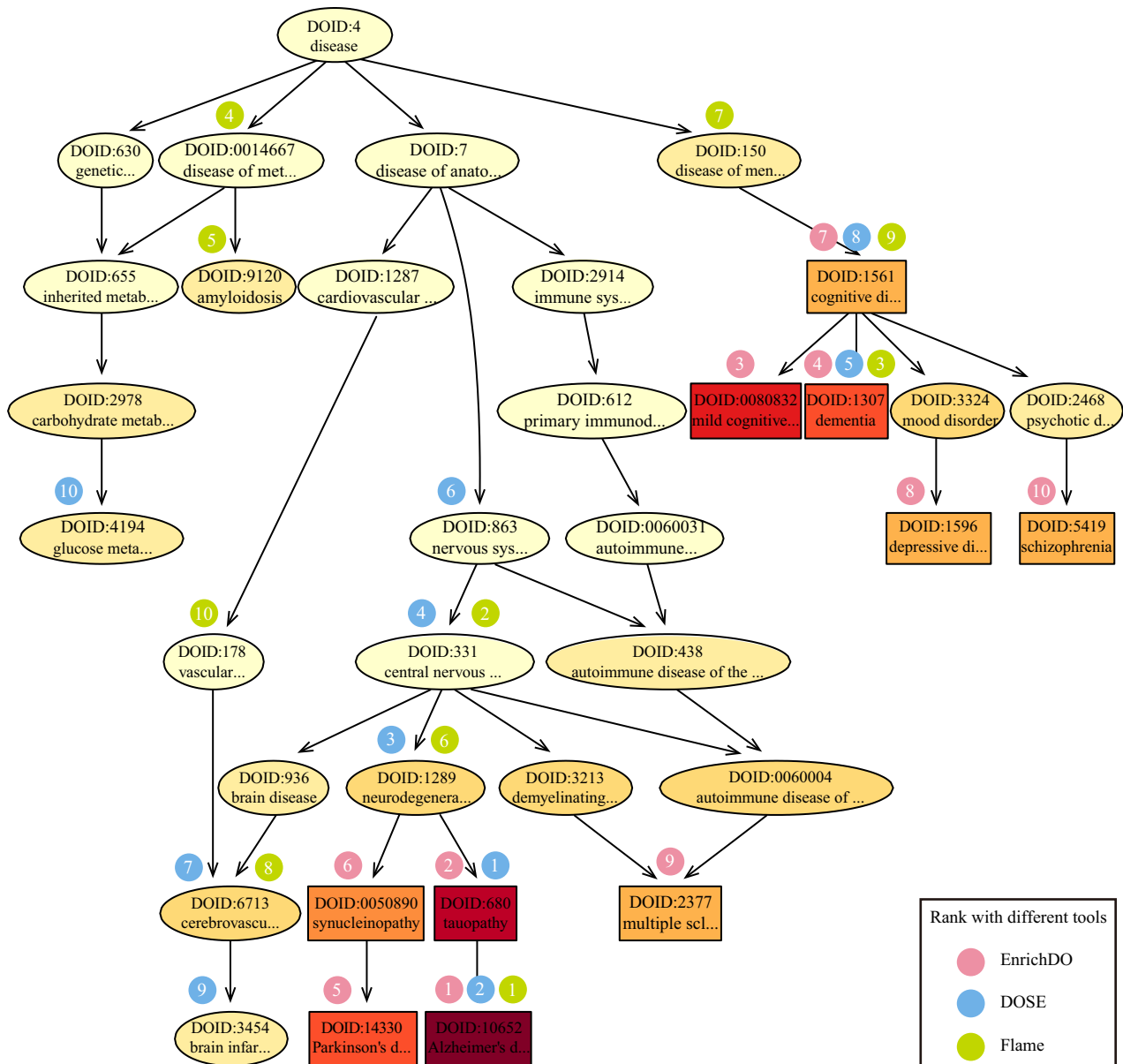


Figure 4: Comparison of enrichment results of EnrichDO with DOSE and Flame on the AD case.

noise from a background gene set, at 5% intervals, repeated 100 times. Significantly overlapped DO terms in the top 100 decreased from 94.8 to 86.6 (Fig. 5B). The results highlight the robustness of EnrichDO to data noise.

Accuracy assessment

The comparison and evaluation of different DO enrichment algorithms relies on true relevant DO terms. In the real datasets mentioned above, the true significant DO terms are not known. To address this problem, the simulated dataset was applied to test the accuracy of EnrichDO. The simulation study was designed as follows:

- (i) *Select the known DO terms.* The number of genes directly annotated with DO terms was distributed from 1 to 4,730. DO terms with annotated genes that are too small or too large are biased. Therefore, we ordered DO terms by the number of their annotated genes and selected 20 (or 30) known DO

terms randomly from the middle 50% of DO terms, which were deemed as the truly enriched nodes.

- (ii) *Obtain the interesting gene list.* The genes annotated to the known DO terms were combined into a single set, which were deemed the list of interesting genes. Considering that some genes annotated with many DO terms are without specificity and may bring noise to simulation study, genes annotated to 30 or more DO terms were removed from the list.
- (iii) *Evaluate performance.* In the simulation study, the performance of the algorithms is evaluated by the number of overlaps between the identified significant DO terms and the known DO terms.

We compared EnrichDO with the hypergeometric test (based on annotations of EnrichDO) and DOSE (Fig. 5C–F). When selecting 20 known DO terms, among top 25 results, EnrichDO identified an average of 9.23 (46.15%) nodes that overlapped with known DO

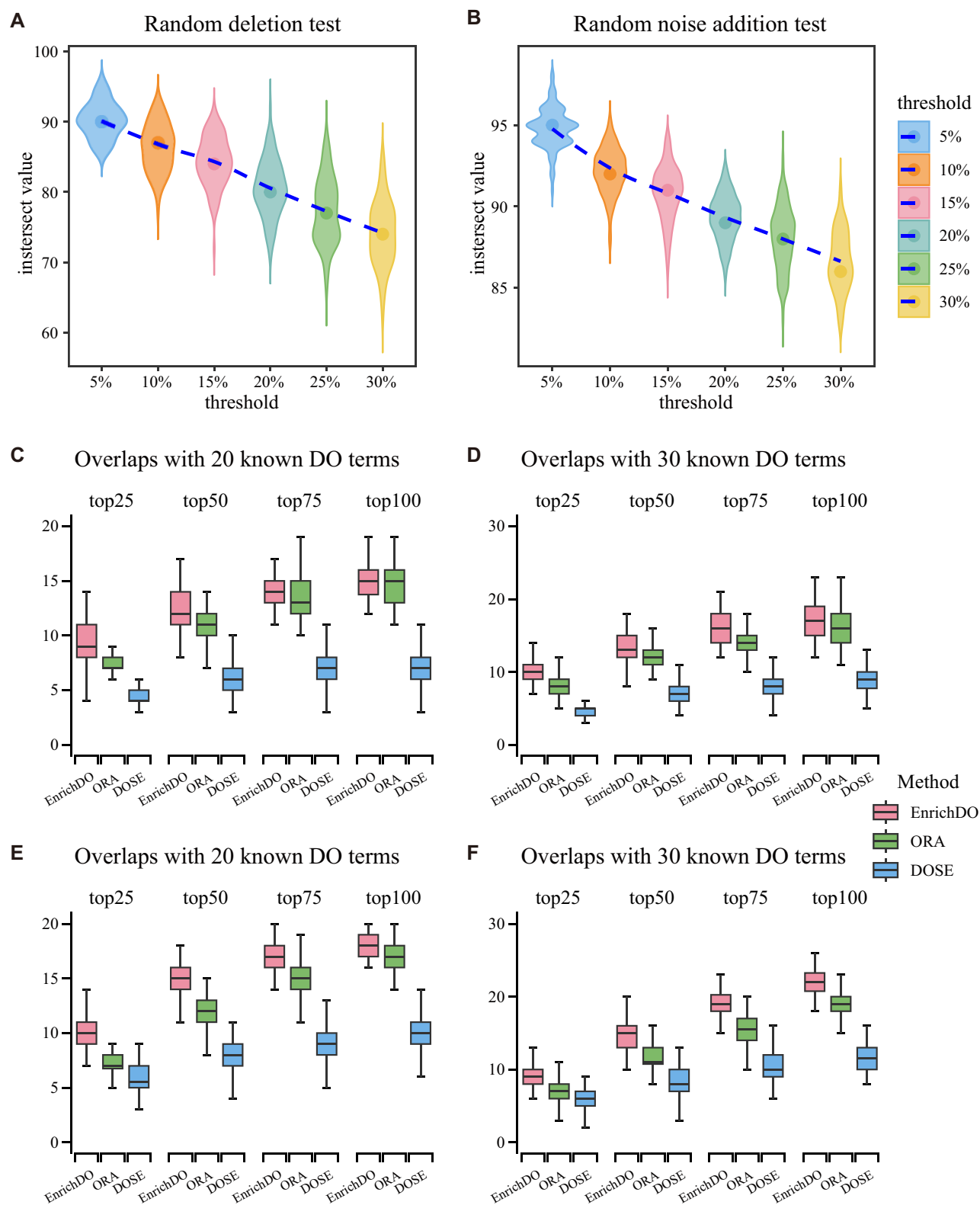


Figure 5: Stability and accuracy assessment of EnrichDO. (A, B) Stability assessment of EnrichDO by random deletion test and noise addition test on the AD case. (C, D) Accuracy assessment of EnrichDO with simulated dataset (semantic expansion genes) of 20 and 30 known DO terms. (E, F) Accuracy assessment of EnrichDO with simulated dataset (directly annotated genes) of 20 and 30 known DO terms.

terms, compared with an average of 7.57 (37.85%) overlaps for the hypergeometric test and 4.39 (21.95%) overlaps for DOSE. When relaxed to the top 100 nodes, an average of 14.79, 14.73, and 7.2 (73.95%, 73.65%, and 36%) overlaps were identified by EnrichDO, the hypergeometric test, and DOSE, respectively (Fig. 5C). The results indicate that EnrichDO has a better performance. A simulation study of 30 selected, known DO terms was implemented in a similar manner; results are shown in Fig. 5D. Notably, when the interesting gene list was derived from genes directly annotated with 20 (or 30) known DO terms, accuracy improved significantly, from 73.95% to 89.7% (or 57% to 72.93%), as shown in Fig. 5E, F. Thus, the accuracy of EnrichDO was better than the hypergeometric test, indicating that the weighted algorithm performs better than the classic ORA enrichment methods. The hypergeometric test was better than DOSE, indicating the accuracy of disease–gene annotations of EnrichDO.

EnrichDO application on disease expression profile dataset

EnrichDO was further applied to 2 real gene expression datasets of pancreatic cancer [49, 50]. The pancreatic cancer dataset I contains microarray gene expression profiles (GSE16515), including 36 pancreatic tumors and 16 normal samples [49]. Briefly, a list of 1,380 differentially expressed genes (DEGs) were identified with a strict threshold ($|\log FC| \geq 1$, $P < 0.05$, $q < 0.05$), of which 1,094 up-regulated DEGs were taken as the interesting genes. DO enrichment results using EnrichDO, DOSE, and the hypergeometric test are displayed in Table 4 and Fig. 6A. EnrichDO identified “pancreatic cancer” (DOID:1793) as the most significantly enriched term, as did the hypergeometric test. Meanwhile, DOSE identified it as the 8th most significantly enriched term and KOBAS as the 21st and 22nd most significantly enriched terms. Among top 10 results, EnrichDO identified many cancer types, such as “stomach cancer” (DOID:76), “colorectal cancer” (DOID:9256), “breast cancer” (DOID:1612), and “lung non-small cell carcinoma” (DOID:3908). On the one hand, the interesting gene list contains many cancer genes, such as *ERBB2*, *HMGA1*, and *MET*, that are related to these cancer types, which have been widely studied. Because these cancer terms have thousands of directly annotated genes, they may be apt to enrichment. Such case also took place in DOSE and the hypergeometric test results. Furthermore, DO terms identified by DOSE and the hypergeometric test were more general—for example, “endocrine gland cancer” (DOID:170), “gastrointestinal system cancer” (DOID:3119), and “carcinoma” (DOID:305), which were ancestors of the results of EnrichDO. The above results indicate that EnrichDO has superior performance over other algorithms in real microarray gene expression datasets.

Pancreatic cancer dataset II contains RNA sequencing expression profiles (GSE119794), including 10 paired pancreatic tumor and normal samples [50]. After differential gene expression analysis, a list of 372 upregulated DEGs were identified with a strict threshold ($|\log FC| \geq 1$, $P < 0.05$, $q < 0.05$), which were adopted as the interesting gene list. As Fig. 6B shows, EnrichDO obtained enrichment results similar to those for dataset I, even though the number of interesting genes varied significantly. Among top 10 results, the terms identified by EnrichDO were more specific than the hypergeometric test, although “pancreatic cancer” ranked second in the EnrichDO results, highlighting EnrichDO’s utility for RNA sequencing expression dataset analysis. In contrast, DOSE identified “pancreatic cancer” as the 44th most significantly enriched term.

The results show that the annotations and proposed algorithms perform better than the classic methods on real gene expression datasets.

EnrichDO application on other datasets

DO enrichment on a microbial host gene set

To explore the application of EnrichDO on additional datasets, we applied EnrichDO to the host gene set of microbes of inflammatory bowel disease (IBD). IBD, mainly in forms of Crohn’s disease (CD) and ulcerative colitis (UC), is characterized by debilitating and chronic relapsing and remitting inflammation of the gastrointestinal tract or the colon; IBD exhibits significant heterogeneity at the clinical, molecular, genetic, and microbial levels [51]. Priya et al. [52] investigated the shared and disease-specific host gene–microbiome associations and identified subsets of significantly correlated host genes and gut microbes. The set of host genes was used to perform DO enrichment analysis in this study. EnrichDO identified IBD (DOID:0050589) accurately, ranking first in the result list, with “colorectal cancer” (DOID:9256) ranking third. The second term was “asthma” (DOID:2841), which was found to be associated with subsequent development of IBD by Kuenzig et al. [53]. As shown in Fig. 6C, the results suggest a high level of accuracy of EnrichDO when using coexpressed host genes in microorganisms, supporting the hypothesis that host genes and gut microbial taxa involved in common biological functions act in a coordinated fashion.

DO enrichment on hallmark gene sets

EnrichDO can also be applied to investigate potential connections between diseases and specific gene sets, such as hallmark gene sets. We selected the hallmark gene sets of “inflammatory response” and “pancreas beta cells” from the Molecular Signatures Database (MSigDB) [54] and used the gene sets to execute DO enrichment analysis. For “inflammatory response,” the top 10 enriched DO terms in EnrichDO were inflammation-related diseases or immune diseases (see Table 5 and Fig. 6D). For “pancreas beta cells,” the top 10 enriched DO terms in EnrichDO consisted entirely of islet-related diseases or diabetes mellitus (see Table 6 and Fig. 6E). The results suggest that EnrichDO is an effective means of uncovering associations between specified gene sets and human diseases.

Discussion

Compared to various GO enrichment analysis methods, DO enrichment analysis methods are relatively scarce. We annotated the latest GeneRIF information with DO terms and provided a new DO enrichment analysis method, EnrichDO, which comprehensively considers DO graph topology and double-weighting of annotated genes. EnrichDO is a moderately weighted algorithm that effectively addresses over-enrichment without pruning truly enriched nodes. EnrichDO improved the accuracy of enrichment analysis results compared to classic enrichment analysis methods and current DO-based enrichment analysis tools. Simulation and robustness tests indicate high and stable performance for EnrichDO. Additionally, we successfully applied EnrichDO to different datasets, such as a microarray gene expression dataset, an RNA sequencing expression dataset, a host gene set of microorganisms, and hallmark gene datasets.

Table 4: Comparison of enrichment analysis results of EnrichDO on pancreatic cancer case (GSE16515)

DOID	DOTerm	Level	geneRatio	EnrichDO		ORA*		DOSE	
				P	Rank	P	Rank	P	Rank
DOID:1793	Pancreatic cancer	6	435/1,038	2.24E-92	1	4.64E-107	1	1.99E-24	8
DOID:10534	Stomach cancer	6	498/1,038	6.82E-90	2	3.67E-90	7	1.82E-16	20
DOID:299	Adenocarcinoma	6	488/1,038	6.17E-89	3	1.04E-100	2	2.42E-10	59
DOID:1324	Lung cancer	6	592/1,038	4.03E-85	4	5.23E-89	9	1.86E-15	24
DOID:850	Lung disease	5	646/1,038	2.83E-80	5	1.20E-96	4	1.47E-12	44
DOID:9256	Colorectal cancer	8	564/1,038	2.34E-79	6	1.09E-79	15	3.43E-09	72
DOID:1612	Breast cancer	6	613/1,038	4.62E-79	7	2.18E-79	18	2.08E-25	6
DOID:3908	Lung non-small cell carcinoma	8	481/1,038	6.27E-79	8	2.68E-81	13	3.15E-14	28
DOID:76	Stomach disease	4	501/1,038	9.22E-79	9	1.11E-89	8	1.10E-03	241
DOID:0050615	Respiratory system cancer	5	615/1,038	1.97E-77	10	9.96E-91	6	5.30E-16	22
DOID:26	Pancreas disease	4	445/1,038	5.78E-68	16	7.90E-100	3	8.21E-03	376
DOID:28	Endocrine system disease	3	617/1,038	1.72E-32	48	3.05E-91	5	2.94E-02	495
DOID:170	Endocrine gland cancer	5	633/1,038	9.90E-50	24	5.71E-87	10	1.66E-27	4
DOID:0050687	Cell type cancer	4	814/1,038	1	4618	1	4674	1.66E-35	1
DOID:0050686	Organ system cancer	4	944/1,038	1	4617	1	4673	7.31E-32	2
DOID:305	Carcinoma	5	792/1,038	1	4497	1	4606	2.23E-30	3
DOID:170	Endocrine gland cancer	5	633/1,038	9.90E-50	24	5.71E-87	10	1.66E-27	4
DOID:3119	Gastrointestinal system cancer	5	770/1,038	1	4499	1	4608	2.44E-26	5
DOID:5093	Thoracic cancer	5	613/1,038	2.54E-71	13	2.40E-79	19	2.38E-25	7
DOID:120	Female reproductive organ cancer	6	493/1,038	3.07E-56	21	6.03E-75	21	2.62E-24	9
DOID:193	Reproductive organ cancer	5	587/1,038	1.14E-44	29	2.40E-71	24	1.00E-23	10

*ORA represents the hypergeometric test.

There are some limitations of the current study. One is that the size of the interesting gene list influences the enrichment analysis results. As shown in the simulation study, when the size of known DO terms increased from 20 to 30, the average number of interesting genes (directly annotated) increased from 1,525 to 2,176, while accuracy decreased from 89.7 to 72.93% (Fig. 5E, F). It can be concluded that an excessively large genes-of-interest set can decrease enrichment efficiency. This phenomenon also appeared in the pancreatic cancer (GSE16515) case. By setting different thresholds of |logFC|, the number of interesting genes changed, and EnrichDO acquired better performance within the number range 67–1,094 (see Supplementary Fig. S1). When the genes-of-interest set is large, some general DO terms with too many annotated genes are enriched; examples include “stomach cancer” and “lung non-small cell carcinoma,” which may be related to pancreatic cancer. It is worth noting that some top-ranked terms, such as “stomach disease” (23/3,141) and “lung disease” (119/4,701), may be over-enriched, owing to inheritance problems, although EnrichDO down-weighted indirectly annotated genes. Therefore, a moderate size for the interesting gene list is suggested, as is setting a threshold for the number of annotated gene of DO terms (e.g., minNum=10, maxNum=2,000). Another unexpected issue we observed is bias of annotations. Some specific DO terms have too few annotated genes to be enriched, due to a lack of available information. Out of 4,361 DO terms, 2,971 were annotated to fewer than 10 genes. A pipeline that automatically integrates the manually curated disease–gene relations from other databases will greatly improve the performance of EnrichDO.

In general, annotations of the human genome with DO terms and the corresponding weighted DO enrichment analysis method described herein demonstrate superior performance over existing approaches. Semantically annotating results based on the latest GeneRIF and DO terms connected data on genes and dis-

eases through the lens of human disease. Based on these annotations, EnrichDO exhibited higher accuracy that often yielded more specific significant DO terms, which alleviated the inheritance problem, making it an effective DO enrichment tool. To facilitate the use of our model, we have developed an R-based software package, which is freely available through Bioconductor [35] or at GitHub [36]. Background annotations will be updated annually according to the latest information in the GeneRIF and DO databases, and the EnrichDO package will be continuously maintained.

Availability of Supporting Source Code and Requirements

Project name: EnrichDO
Project homepage: <https://bioconductor.org/packages/release/bioc/html/EnrichDO.html> [35], the latest version can be available at Bioconductor devel branch [55] or GitHub [36]. Our GitHub project has also been archived in Software Heritage [56].
Software Heritage PID: swh:1:snp:6614022663955ab147f83f12ecc721ea8f314b56
Operating system(s): Platform independent
Programming language: It is recommended to use R version 4.4 and Bioconductor version 3.20 or later. For versions below R-4.4 (minimum R-4.0), users can download source packages from Bioconductor and install package(s) from local files.
Other requirements: R packages BiocGenerics, Rgraphviz, clusterProfiler, hash, S4Vectors, dplyr, ggplot2, graph, magrittr, methods, pheatmap, graphics, utils, purrr, tidy, stats
License: MIT
RRID:SCR_025840
bioTools ID: biotools:enrichdo
This workflow is also available in WorkflowHub [57].

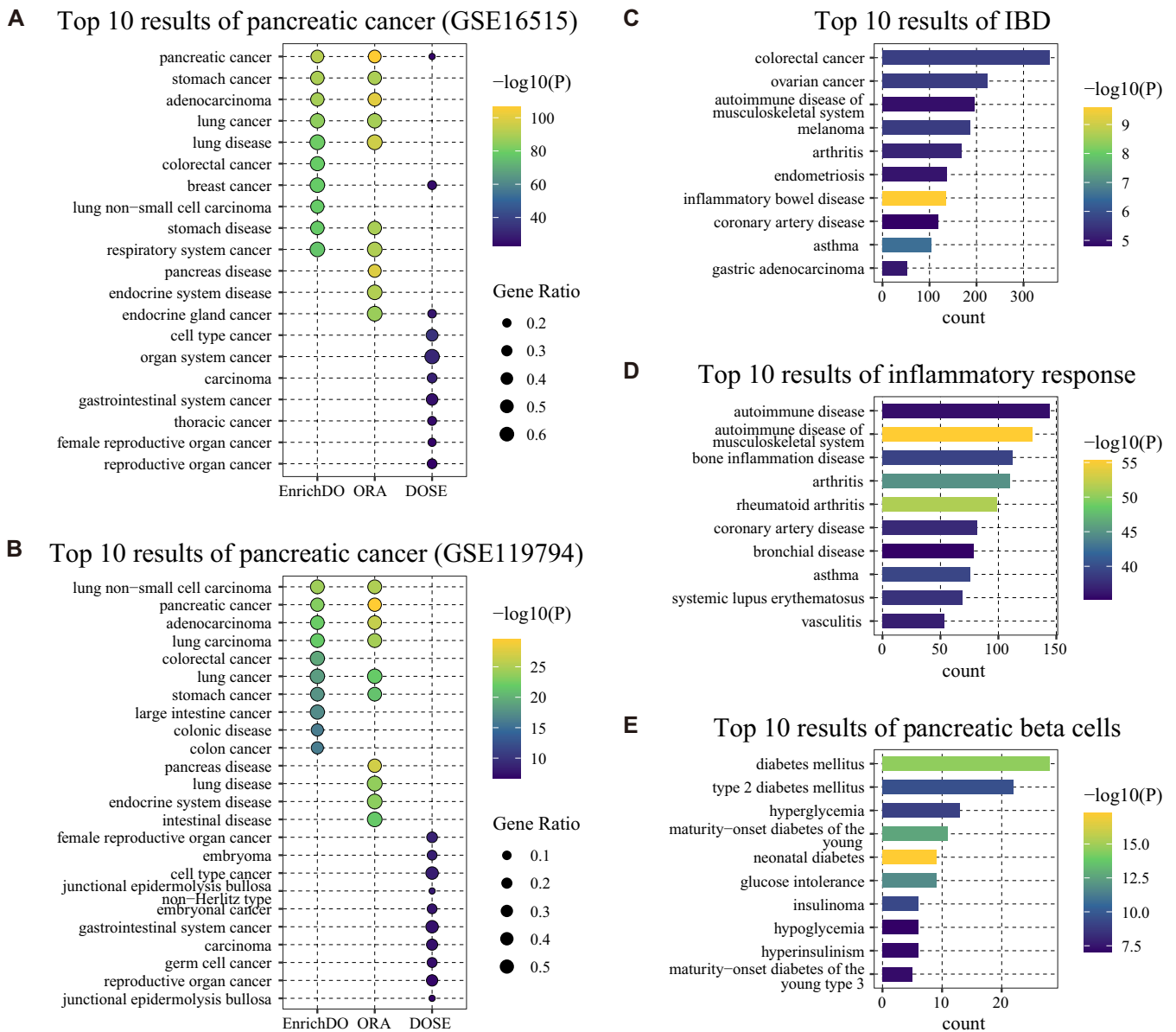


Figure 6: Enrichment results of various datasets. (A) Top 10 enrichment results of EnrichDO, ORA (hypergeometric test), and DOSE for microarray gene expression profiles (GSE16515). (B) Top 10 enrichment results of EnrichDO, ORA (hypergeometric test), and DOSE for RNA sequencing expression profiles (GSE119794). (C) Top 10 enrichment results of EnrichDO for host genes of microbes of IBD. (D) Top 10 enrichment results of EnrichDO for hallmark gene sets of “inflammatory response.” (E) Top 10 enrichment results of EnrichDO for hallmark gene sets of “pancreas beta cells.”

Table 5: The top 10 enrichment analysis results of EnrichDO on an inflammatory response case

DOID	DOTerm	P	P-adjust	geneRatio	bgRatio
DOID:0060032	Autoimmune disease of musculoskeletal system	5.04E-56	2.43E-52	129/200	2,083/15,106
DOID:7148	Rheumatoid arthritis	4.59E-52	1.11E-48	99/200	1,235/15,106
DOID:848	Arthritis	1.43E-45	2.30E-42	110/200	1,735/15,106
DOID:2841	Asthma	2.96E-40	3.56E-37	76/200	905/15,106
DOID:3342	Bone inflammation disease	4.70E-40	4.52E-37	112/200	1,819/15,106
DOID:9074	Systemic lupus erythematosus	8.41E-39	6.75E-36	69/200	780/15,106
DOID:3393	Coronary artery disease	2.96E-38	2.03E-35	82/200	1,161/15,106
DOID:865	Vasculitis	1.95E-37	1.17E-34	53/200	385/15,106
DOID:417	Autoimmune disease	1.64E-36	8.76E-34	144/200	2,877/15,106
DOID:1176	Bronchial disease	5.42E-36	2.61E-33	79/200	952/15,106

Table 6: The top 10 enrichment analysis results of EnrichDO on a pancreas beta cells case.

DOID	DOTerm	P	P-adjust	geneRatio	bgRatio
DOID:11717	Neonatal diabetes	5.71E-18	2.75E-14	9/39	27/15,106
DOID:9351	Diabetes mellitus	2.75E-15	6.61E-12	28/39	2,087/15,106
DOID:0050524	Maturity-onset diabetes of the young	2.59E-13	4.15E-10	11/39	101/15,106
DOID:10603	Glucose intolerance	1.90E-12	2.28E-09	9/39	95/15,106
DOID:9352	Type 2 diabetes mellitus	2.44E-10	2.35E-07	22/39	1,449/15,106
DOID:3892	Insulinoma	5.14E-10	4.13E-07	6/39	38/15,106
DOID:4195	Hyperglycemia	1.01E-09	6.92E-07	13/39	324/15,106
DOID:0111102	Maturity-onset diabetes of the young type 3	5.40E-08	3.25E-05	5/39	7/15,106
DOID:2018	Hyperinsulinism	6.95E-08	3.72E-05	6/39	84/15,106
DOID:9993	Hypoglycemia	9.20E-08	4.43E-05	6/39	89/15,106

Additional Files

Supplementary Fig. S1. Statistics of enrichment results according to different threshold of |logFC|.

Supplementary Materials. Comparison with Other Methods.

Abbreviations

AD: Alzheimer’s disease; ALL: acute lymphoblastic leukemia; BP: biological process; CD: Crohn’s disease; ChEBI: Chemical Entities of Biological Interest; DAG: directed acyclic graph; DEG: differentially expressed gene; DGA: Disease and Gene Annotations database; DO: Disease Ontology; GeneRIF: Gene Reference Into Function; GO: Gene Ontology; HPO: Human Phenotype Ontology; IBD: inflammatory bowel disease; IL6: interleukin 6; MCI: mild cognitive impairment; MMP9: matrix metalloproteinase 9; MMTx: MetaMap Transfer tool; MSigDB: Molecular Signatures Database; NHGRI: NHGRI GWAS catalog; OAE: Ontology of Adverse Events; ORA: overrepresentation analysis; TGFB1: transforming growth factor beta 1; TNF: tumor necrosis factor; TP53: tumor protein p53; UC: ulcerative colitis; UMLS: Unified Medical Language System; VEGFA: vascular endothelial growth factor A.

Author Contributions

Liang Cheng (Design [lead], Funding acquisition [lead], Writing—review & editing [equal]), Haixiu Yang (Data curation [lead], Funding acquisition [equal], Methodology [lead], Software [equal], Writing—original draft [lead]), Hongyu Fu (Software [lead], Validation [equal], Visualization [lead]), Meiyi Zhang (Formal analysis [lead], Validation [lead]), Yangyang Liu (Formal analysis [equal]), Yongqun Oliver He (Writing—review & editing [equal]), and Chao Wang (Visualization [equal]).

Funding

This work was supported by National Natural Science Foundation of China (grant numbers 62222104, 62172130 to L.C. and 61902095 to H.X.Y.), and Heilongjiang Postdoctoral Fund (LBH-Q20030 to L.C.).

Data Availability

Annotations of the Human Genome with DO were stored in <https://github.com/liangcheng-hrbmu/EnrichDO/blob/devel/data/doterminals.rda>.

The datasets for case studies of enrichment analysis were collected as follows:

- The dataset of the ALL case [45] was obtained from the R package from Bioconductor [58], and the differentially expressed genes were extracted.
- The dataset of the AD case was curated using genes downloaded from DisGeNET [42].
- The expression profile datasets were downloaded from the Gene Expression Omnibus (GEO) database, namely GSE16515 and GSE119794 [49, 50], and the differentially expressed genes were extracted.
- The shared and disease-specific host gene–microbiome associations were obtained from the study of Priya [52].
- The hallmark gene sets were downloaded from the Human Molecular Signatures Database [54] (MSigDB, <https://www.gsea-msigdb.org/gsea/msigdb/>), including HALLMARK_PANCREAS_BETA_CELLS and HALLMARK_INFLAMMATORY_RESPONSE.
- Specific data information has been uploaded to <https://github.com/liangcheng-hrbmu/EnrichDO/tree/devel/thesisData>.

Competing Interests

The authors declare that they have no competing interests.

References

1. Alterovitz G, Xiang M, Hill DP, et al. Ontology engineering. *Nat Biotechnol.* 2010;28(2):128–30. <https://doi.org/10.1038/nbt0210-128>.

2. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000;25(1):25–29. <https://doi.org/10.1038/75556>.

3. Gene Ontology C, Aleksander SA, Balhoff J, et al. The gene ontology knowledgebase in 2023. *Genetics.* 2023;224(1):iyad031. <https://doi.org/10.1093/genetics/iyad031>.

4. Schriml LM, Arze C, Nadendla S, et al. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res.* 2012;40(Database issue):D940–46. <https://doi.org/10.1093/nar/gkr972>.

5. Baron JA, Johnson CS, Schor MA, et al. The DO-KB Knowledgebase: a 20-year journey developing the disease open science ecosystem. *Nucleic Acids Res.* 2024;52(D1):D1305–D14. <https://doi.org/10.1093/nar/gkad1051>.

6. Robinson PN, Kohler S, Bauer S, et al. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet.* 2008;83(5):610–15. <https://doi.org/10.1016/j.ajhg.2008.09.017>.

7. Gargano MA, Matentzoglou N, Coleman B, et al. The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic Acids Res.* 2024;52(D1):D1333–46. <https://doi.org/10.1093/nar/gkad1005>.
8. Degtyarenko K, de Matos P, Ennis M, et al. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* 2008;36(Database issue):D344–50. <https://doi.org/10.1093/nar/gkm791>.
9. Hastings J, Owen G, Dekker A, et al. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.* 2016;44(D1):D1214–19. <https://doi.org/10.1093/nar/gkv1031>.
10. He Y, Sarntinivajai S, Lin Y, et al. OAE: the ontology of adverse events. *J Biomed Semantics.* 2014;5:29. <https://doi.org/10.1186/2041-1480-5-29>.
11. Noy NF, Shah NH, Whetzel PL, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* 2009;37(Web Server issue):W170–73. <https://doi.org/10.1093/nar/gkp440>.
12. Whetzel PL, Noy NF, Shah NH, et al. BioPortal: enhanced functionality via new web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.* 2011;39(Web Server issue):W541–45. <https://doi.org/10.1093/nar/gkr469>.
13. Smith B, Ashburner M, Rosse C, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol.* 2007;25(11):1251–55. <https://doi.org/10.1038/nbt1346>.
14. Jackson R, Matentzoglou N, Overton JA, et al. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. 2021. Database (Oxford). <https://doi.org/10.1093/database/baab069>. Accessed 28 March 2024.
15. Carbon S, Ireland A, Mungall CJ, et al. AmiGO: online access to ontology and annotation data. *Bioinformatics.* 2009;25(2):288–89. <https://doi.org/10.1093/bioinformatics/btn615>.
16. Thomas PD, Hill DP, Mi H, et al. Gene ontology causal activity modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. *Nat Genet.* 2019;51(10):1429–33. <https://doi.org/10.1038/s41588-019-0500-1>.
17. Yu G, Li F, Qin Y, et al. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics.* 2010;26(7):976–78. <https://doi.org/10.1093/bioinformatics/btq064>.
18. Thomas PD, Ebert D, Muruganujan A, et al. PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci.* 2022;31(1):8–22. <https://doi.org/10.1002/pro.4218>.
19. Zemojtel T, Kohler S, Mackenroth L, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med.* 2014;6(252):252ra123. <https://doi.org/10.1126/scitranslmed.3009262>.
20. Li J, Gong B, Chen X, et al. DOSim: an R package for similarity between diseases based on disease ontology. *BMC Bioinform.* 2011;12:266. <https://doi.org/10.1186/1471-2105-12-266>.
21. Cheng L, Li J, Ju P, et al. SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. *PLoS One.* 2014;9(6):e99415. <https://doi.org/10.1371/journal.pone.0099415>.
22. Hu Y, Zhao L, Liu Z, et al. DisSetSim: an online system for calculating similarity between disease sets. *J Biomed Semantics.* 2017;8(Suppl 1):28. <https://doi.org/10.1186/s13326-017-0140-2>.
23. Osborne JD, Flatow J, Holko M, et al. Annotating the human genome with disease ontology. *BMC Genomics.* 2009;10(Suppl 1):S6. <https://doi.org/10.1186/1471-2164-10-s1-s6>.
24. LePendu P, Musen MA, Shah NH. Enabling enrichment analysis with the human disease ontology. *J Biomed Inform.* 2011;44(Suppl 1):S31–38. <https://doi.org/10.1016/j.jbi.2011.04.007>.
25. Peng K, Xu W, Zheng J, et al. The Disease and Gene Annotations (DGA): an annotation resource for human disease. *Nucleic Acids Res.* 2013;41(Database issue):D553–60. <https://doi.org/10.1093/nar/gks1244>.
26. Bu D, Luo H, Huo P, et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.* 2021;49(W1):W317–25. <https://doi.org/10.1093/nar/gkab447>.
27. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 2016;44(W1):W90–97. <https://doi.org/10.1093/nar/gkw377>.
28. Elizarraras JM, Liao Y, Shi Z, et al. WebGestalt 2024: faster gene set analysis and new support for metabolomics and multi-omics. *Nucleic Acids Res.* 2024;52(W1):W415–21. <https://doi.org/10.1093/nar/gkae456>.
29. Karatzas E, Baltoumas FA, Aplakidou E, et al. Flame (v2.0): advanced integration and interpretation of functional enrichment results from multiple sources. *Bioinformatics.* 2023;39(8):btad490. <https://doi.org/10.1093/bioinformatics/btad490>.
30. Scholz C, Lyon D, Refsgaard JC, et al. Avoiding abundance bias in the functional annotation of post-translationally modified proteins. *Nat Methods.* 2015;12(11):1003–4. <https://doi.org/10.1038/nmeth.3621>.
31. Yu G, Wang LG, Yan GR, et al. DOSE: an R/bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics.* 2015;31(4):608–9. <https://doi.org/10.1093/bioinformatics/btu684>.
32. Amar D, Vitzel A, Levy C, et al. ADEPTUS: a discovery tool for disease prediction, enrichment and network analysis based on profiles from many diseases. *Bioinformatics.* 2018;34(11):1959–61. <https://doi.org/10.1093/bioinformatics/bty027>.
33. Alexa A, Rahnenfuhrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics.* 2006;22(13):1600–7. <https://doi.org/10.1093/bioinformatics/btl140>.
34. Grossmann S, Bauer S, Robinson PN, et al. Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. *Bioinformatics.* 2007;23(22):3024–31. <https://doi.org/10.1093/bioinformatics/btm440>.
35. Bioconductor. EnrichDO: a Global Weighted Model for Disease Ontology Enrichment Analysis (version 1.0). 2024. <https://doi.org/10.18129/B9.bioc.EnrichDO>. Accessed 29 October 2024.
36. EnrichDO. <https://github.com/liangcheng-hrbmu/EnrichDO>. Accessed 7 February 2025.
37. Schriml LM, Munro JB, Schor M, et al. The Human Disease Ontology 2022 update. *Nucleic Acids Res.* 2022;50(D1):D1255–61. <https://doi.org/10.1093/nar/gkab1063>.
38. GeneRIF: Gene Reference into Function. <https://ftp.ncbi.nih.gov/gene/GeneRIF/>. Accessed 26 August 2024.
39. Dai M, Shah NH, Xuan W, et al. An efficient solution for mapping free text to ontology terms. *AMIA Summit on Translational Bioinformatics.* 2008;21.
40. Shah NH, Bhatia N, Jonquet C, et al. Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC*

- Bioinf. 2009;10(Suppl 9):S14. <https://doi.org/10.1186/1471-2105-10-S9-S14>.
41. Annibaldi A, Meier P. Checkpoints in TNF-induced cell death: implications in inflammation and cancer. *Trends Mol Med.* 2018;24(1):49–65. <https://doi.org/10.1016/j.molmed.2017.11.002>.
 42. Pinero J, Ramirez-Angueta JM, Sauch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 2020;48(D1):D845–55. <https://doi.org/10.1093/nar/gkz1021>.
 43. Rappaport N, Twik M, Plaschkes I, et al. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res.* 2017;45(D1):D877–87. <https://doi.org/10.1093/nar/gkw1012>.
 44. Grissa D, Junge A, Oprea TI, et al. Diseases 2.0: a weekly updated database of disease-gene associations from text mining and data integration. *Database (Oxford).* 2022. <https://doi.org/10.1093/database/baac019>. Accessed 8 April 2024.
 45. Chiaretti S, Li X, Gentleman R, et al. Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood.* 2004;103(7):2771–78. <https://doi.org/10.1182/blood-2003-09-3243>.
 46. Ballard C, Gauthier S, Corbett A, et al. Alzheimer's disease. *Lancet.* 2011;377(9770):1019–31. [https://doi.org/10.1016/s0140-6736\(10\)61349-9](https://doi.org/10.1016/s0140-6736(10)61349-9).
 47. Horie K, Barthélemy NR, Spina S, et al. CSF tau microtubule-binding region identifies pathological changes in primary tauopathies. *Nat Med.* 2022;28(12):2547–54. <https://doi.org/10.1038/s41591-022-02075-9>.
 48. Reitz C, Mayeux R. Alzheimer disease: epidemiology, diagnostic criteria, risk factors and biomarkers. *Biochem Pharmacol.* 2014;88(4):640–51. <https://doi.org/10.1016/j.bcp.2013.12.024>.
 49. Pei H, Li L, Fridley BL, et al. FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt. *Cancer Cell.* 2009;16(3):259–66. <https://doi.org/10.1016/j.ccr.2009.07.016>.
 50. Lin J, Wu YJ, Liang X, et al. Network-based integration of mRNA and miRNA profiles reveals new target genes involved in pancreatic cancer. *Mol Carcinog.* 2019;58(2):206–18. <https://doi.org/10.1002/mc.22920>.
 51. Adolph TE, Zhang J. Diet fuelling inflammatory bowel diseases: preclinical and clinical concepts. *Gut.* 2022;71(12):2574–86. <https://doi.org/10.1136/gutjnl-2021-326575>.
 52. Priya S, Burns MB, Ward T, et al. Identification of shared and disease-specific host gene-microbiome associations across human diseases using multi-omic integration. *Nat Microbiol.* 2022;7(6):780–95. <https://doi.org/10.1038/s41564-022-01121-z>.
 53. Kuenzig ME, Barnabe C, Seow CH, et al. Asthma is associated with subsequent development of inflammatory bowel disease: a population-based case-control study. *Clin Gastroenterol Hepatol.* 2017;15(9):1405–12. e3. <https://doi.org/10.1016/j.cgh.2017.02.042>.
 54. Castanza AS, Recla JM, Eby D, et al. Extending support for mouse data in the Molecular Signatures Database (MSigDB). *Nat Methods.* 2023;20(11):1619–20. <https://doi.org/10.1038/s41592-023-0147-7>.
 55. Bioconductor. EnrichDO (development version): a Global Weighted Model for Disease Ontology Enrichment Analysis (version 1.1.1). 2024. Accessed 21 January 2025. <https://bioconductor.org/packages/devel/bioc/html/EnrichDO.html>
 56. Yang H, Fu H, Zhang M, et al. EnrichDO: a global weighted model for disease ontology enrichment analysis (Version 1) [Computer software]. *Software Heritage.* 2025. https://archive.softwareheritage.org/browse/snapshot/6614022663955ab147f83f12ecc721ea8f314b56/directory/?origin_url=https://github.com/liangcheng-hrbmu/EnrichDO. Accessed 21 January 2025.
 57. Cheng L. EnrichDO: a global weighted model for disease ontology enrichment analysis. *WorkflowHub.* 2024. <https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.1221.1>. Accessed 21 January 2025.
 58. Li X. ALL: A data package. R package version 1.48.0. *Bioconductor.* 2024. <https://doi.org/doi:10.18129/B9.bioc.ALL>