Resource

Gene Set to Diseases (GS2D): Disease Enrichment Analysis on Human Gene Sets with Literature Data

Miguel A. Andrade-Navarro[1,2], Jean Fred Fontaine[1,2,*]

[1] Faculty of Biology, Johannes Gutenberg University of Mainz, 55128 Mainz, Germany
[2] Institute of Molecular Biology, 55128 Mainz, Germany

[*] Correspondence: fontaine@uni-mainz.de

Received 2016-05-12; Accepted 2016-10-16; Published 2016-10-30

ABSTRACT

Large sets of candidate genes derived from high-throughput biological experiments can be characterized by functional enrichment analysis. The analysis consists of comparing the functions of one gene set against that of a background gene set. Then, functions related to a significant number of genes in the gene set are expected to be relevant. Web tools offering disease enrichment analysis on gene sets are often based on gene-disease associations from manually curated or experimental data that is accurate but does not cover all diseases discussed in the literature. Using associations automatically derived from literature data could be a cost effective method to improve the coverage of diseases for enrichment analysis at comparable levels of accuracy.

We implemented a method named Gene set to Diseases, GS2D, as a web tool performing disease enrichment analysis on human protein coding gene sets. It uses an automatically built dataset of more than 63 thousand gene-disease associations defined as statistically significant co-occurrences of genes and diseases in annotations of biomedical citations from PubMed. The dataset covers more diseases for enrichment analysis than the largest comparable curated database (Comparative Toxicogenomics Database) and its performance compared favourably to similar approaches based on manually curated or experimental data. Graphical and programmatic interfaces are available at http://cbdm.uni-mainz.de/geneset2diseases.

KEYWORDS

Bioinformatics; Enrichment analysis; Genes; Diseases; Literature analysis

AVAILABILITY AND REQUIREMENTS

- Project name: Gene set to Diseases
- http://cbdm.uni-mainz.de/geneset2diseases
- Interfaces: HTML/Javascript and REST
- License: free for academic users

INTRODUCTION

Functional enrichment analysis is performed to characterize gene sets derived from high-throughput technologies such as next generation sequencing. The analysis compares the functional annotations of one gene set against those of a background gene set to find functions significantly enriched in the selected gene set. These functional annotations can be retrieved from different databases such as gene databases (e.g. Gene Ontology [1] terms for molecular functions), biological pathway databases (e.g. WikiPathways [2]), or large collections of experimental datasets (e.g. ENCODE project [3]). If a significant number of genes are involved in the same function then this function can be considered as more likely to be relevant to the experimental conditions related to the gene set. When annotations regarding gene functions are known only from biological experiments that do not cover all possible genes and functions [4], functional enrichment analyses can fail to return all relevant results. Yet, use of computational algorithms can help infer gene functions for more genes (e.g. [5, 6]).

Whereas many tools exist that perform such analyses based on Gene Ontology terms (see the Gene Ontology Consortium pages for a list [1]), only few tools analyse diseases (e.g. ToppGene [7] HPOSim [8], and DOSE [9]). This is in stark contrast to the fact that the study of human disease is a critical focus of many biomedical researchers. These tools often define associations between genes and diseases from curated information (e.g. from the Online Mendelian Inheritance in Man [1] (OMIM®) database, the Comparative Toxicogenomics Database (CTD) [10], or WikiPathways [2]), or experimental datasets such as those from genome-wide association studies (e.g. GWAS Catalog [11]). As there are many diseases for which few or no genes have been associated by such ways, these tools are consequently limited.

The PubMed database which contains more than 26 million citations for biomedical literature (e.g. journal articles) offers an alternative source of data about the relations of genes to diseases. Automated text mining of these data has been already applied to derive gene-disease associations (e.g. [12, 13]); however, these approaches require recognizing gene names using automated text mining methods that suffer from low accuracy [14, 15]. Another way to extract gene-disease associations from the literature is to integrate the numerous manually curated annotations of PubMed citations for genes or diseases.

Our method, called Gene set to Diseases (GS2D), derives gene-disease associations using statistically

significant co-occurrences of genes and diseases in annotations of PubMed citations. GS2D was implemented as a web tool with a graphical and a programmatic interface. Disease enrichment analyses performed using these associations were compared to analyses performed by other approaches.

METHODS

Implementation

The data used as input by GS2D is based on (i) biomedical citations with English abstracts from the PubMed database, (ii) disease terms from the branch C of the Medical Subject Headings thesaurus (MeSH®), and (iii) human protein-coding gene information from the NCBI Gene database [16]. Manual annotations of citations by diseases were extracted from PubMed. Manual annotations of citations by genes were extracted from gene2pubmed and GeneRIF files (NCBI Entrez Gene FTP site). All the data was downloaded, processed and stored in a MySQL database on 16 November 2015.

All computations were limited to 282749 PubMed citations annotated with at least one disease and at least one human protein-coding gene (Figure 1). From the data, gene-disease associations were computed from statistically significant co-occurrences of genes and diseases in annotations of PubMed citations. The significance of each association was evaluated by a p-value from a one-tailed Fisher's exact test and false discovery rate (FDR) calculated by the Benjamini and Hochberg method [17] using the R statistical environment [18]. For a gene G, a disease D, and the literature annotations L, the test is based on a contingency matrix containing the following numbers of citations (Figure 1):

- (a) citations in L annotated with D and G
- (b) citations in L annotated with D but not with G
- (c) citations in L annotated not with D but with G
- (d) citations in L annotated not with D and not with G

Gene-disease associations with less than 3 co-occurrences or a FDR greater than 0.05 were filtered out to produce 63503 associations involving 2214 diseases and 7597 genes.

In order to produce the list of enriched diseases for an input gene set, the diseases associated to the input gene set are compared to the diseases associated to a background gene set. By default, GS2D defines the background gene set as all the human protein-coding genes excluding the input gene set. Significance is evaluated by a p-value from a one-tailed Fisher's exact test (using the R statistical environment) and FDRs calculated by the Benjamini and Hochberg method. For a gene set S, a disease D and a background gene set A, the test is based on a contingency matrix containing the following numbers of genes (Figure 1):

- (a) genes associated with D in S
- (b) genes associated with D not in S (but in A)
- (c) genes not associated with D in S
- (d) genes not associated with D not in S (but in A)

Web pages were built with WordPress 4.4 or programmed using HTML 4, JavaScript, PHP and Perl 5. Data were stored in a MySQL 5.5 database. Web pages were tested using several web browsers (i.e. Firefox 39 and 43, Chrome 47, Chromium 47, Internet Explorer 8 and 11) and operating systems (i.e. Ubuntu 14.04, Windows XP and 8.1).

Benchmarks and comparison

Two types of gene sets were used as gold standards (Figure 2A): 10 GWAS-related gene sets downloaded from the GWAS Catalog resource on 25 November 2015 [11] and 10 non-GWAS-related gene sets of at least 20 genes from Menche et al. [19] who combined OMIM and UniProtKB/Swiss-Prot. As GS2D uses statistical methods similar to those of comparable tools (Fisher's exact test and correction for multiple tests), we benchmarked against the ToppGene web tool [7] that uses different background gene sets derived from manually curated (CTD and OMIM) and experimental data (GWAS datasets) (Figure 2B).

Analyses with GS2D based on automatically derived data from the literature were performed with gene-disease associations defined with at least 5 co-occurrences, at least 2 genes significantly associated with a disease, and FDR$<$0.05. ToppGene was queried to use CTD (ToppGene-CTD), OMIM (ToppGene-OMIM) and GWAS (ToppGene-GWAS) at FDR$<$0.05.

For comparison, CTD manually curated gene-disease associations were downloaded on 4 December 2015. The following criteria were used to select the CTD data: data with direct evidence (not inferred), and data related to human protein-coding genes.

RESULTS

Function A: gene set to disease

GS2D is implemented as a web server to perform disease enrichment analysis on gene sets. The input gene set can be defined with human protein-coding gene symbols or Entrez Gene identifiers. Whereas the popular Gene Ontology enrichment analysis for gene sets is based on gene-function associations, GS2D is based on gene-disease associations inferred by statistically significant co-occurrences of curated annotations of genes and diseases in PubMed citations. GS2D also gives the users options to change the stringency of the results: gene-disease associations can be filtered by a minimum number of co-occurrences (default = 5), and enriched diseases for a gene set can be filtered by FDR (default cutoff = 0.05) or by a minimum number of associated genes from the gene set (default = 2). Enriched diseases are listed with URL links to disease definitions (MeSH web site) and to PubMed citations with co-occurrences (PubMed web site).

Function B: gene to disease / disease to gene

GS2D can also be used to query its associations. In this case, all gene-disease associations related to each input gene or disease are listed with URL links to disease definitions (MeSH web site), gene information (Entrez gene web site) and to PubMed citations with

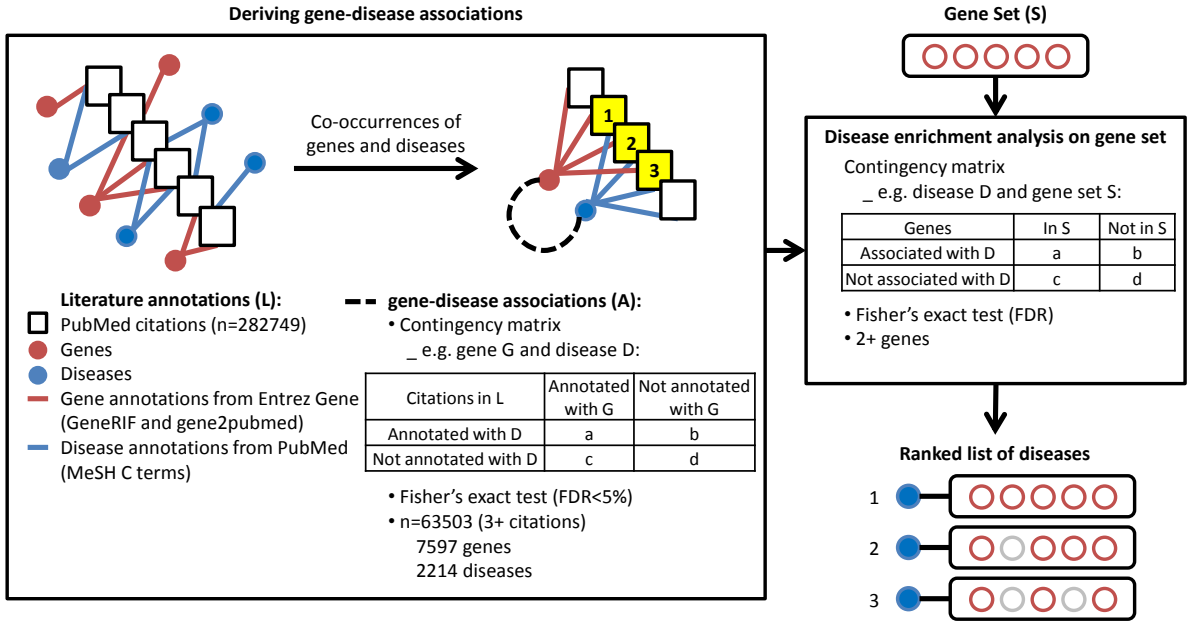


Figure 1: Workflow. Gene set to diseases (GS2D) first derives gene-disease associations (A) from a co-occurrence analysis in the literature annotations (L). These associations are then used to find enriched diseases in a gene set (S).

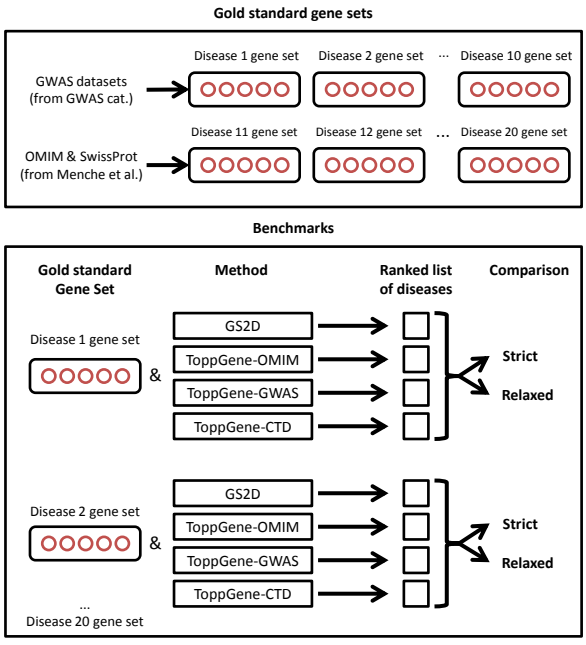


Figure 2: Benchmarks. For benchmarking, 20 gold standard gene sets were selected (10 derived from GWAS Catalog and 10 from Menche et al.). Each gene set was used as input for Gene set to diseases (GS2D), or ToppGene service using gene-disease associations derived from the Comparative Toxicogenomics Database (ToppGene-CTD), the Genome-Wide Association Studies (ToppGene-GWAS), or the Online Mendelian Inheritance in Man (ToppGene-OMIM) database. Results were compared in a strict or relaxed evaluation.

co-occurrences (PubMed web site). A button allows sending all the listed genes as a gene set for a disease enrichment analysis (function A).

Programmatic access

GS2D has a programmatic access allowing batch queries and retrieval of results as tab-separated values. The HTTP-based RESTful API is documented on the web server with examples.

Use case 1: parameter tuning for the analysis of Alzheimer's disease genes

To demonstrate the impact of input parameters on the results, we selected a gene set of 50 human protein-coding genes associated with Alzheimer's disease using the Génie web tool [20]. Génie uses automated document classification to find genes associated with a topic, such as a disease, and it was previously benchmarked for finding Alzheimer's disease genes at a precision of 94% in its top 50 results.

In order to find if there were other diseases involved in a significant fraction of these 50 genes, function A (gene set to disease) of GS2D was used with default parameters. It returned a list of 60 diseases enriched for the gene set, which can be described by count and percentage of associated genes from the gene set, fold change, P-value, FDR and related citations (Figure 3). As expected, the top hit (lowest P-value) was Alzheimer's Disease (42 genes; FDR=4.90e-40).

Evaluating the statistics from the list could help in refining the search to focus on the strongest enrichments. For example, changing the FDR cut-off

to 0.0001 reduced the list from 60 to 21 more confidently enriched diseases. Setting additionally to 10 the minimum number of required co-occurrences for gene-disease associations reduced the list further to only 12 more confidently enriched diseases. It filtered out weaker statistical enrichments such as Creutzfeldt-Jakob Syndrome, which was listed as enriched because of three genes including two with less than ten relevant citations. Notably, the selected diseases were all related to neurodegeneration. This reflects the fact that many genes implicated in Alzheimer's disease have neuronal functions and mutations involved in other neurodegenerative diseases, making it difficult to find biomarker profiles able to distinguish some of these diseases (see for example [21]).

Use case 2: recapitulating p53 involvement in diseases by analysing interacting genes

The second use case involves 254 human genes interacting with p53, previously used as a predictive model for cancer therapy [22]. Knowing that the tumour suppressor p53 is involved in many human cancers [23], we would expect GS2D to recapitulate this knowledge when analysing p53 interacting genes by returning a list of enriched diseases that includes various types of cancers.

In order to get a list of diseases related to each of the 254 input genes, function B (gene to disease) of GS2D was used with default parameters. This resulted in a list of 2638 gene-disease associations that can be described by related citations, fold change, P-value and FDR (Supplementary Table 1). The diversity of diseases that can be associated by literature co-occurrences to a single gene was exemplified by the ABCB1 gene associated with very different diseases including Acute Coronary Syndrome (6 citations), Breast Neoplasms (105 citations) and Epilepsy (53 citations). Sorting results by disease names (by clicking on corresponding column header) allowed visual identification of several diseases associated to multiple genes, including many types of cancer-related diseases such as Adenocarcinoma (42 genes), Adenoma (9 genes), Breast Neoplasms (85 genes) and Cerebellar Neoplasms (2 genes).

In order to know if the multiple associations observed above were significant for the gene set as a whole, function A (gene set to disease) of GS2D was used. It returned a list of 101 diseases significantly enriched for the gene set (Supplementary Table 2). The list included Adenocarcinoma (FDR=4.38e-03), and Breast Neoplasms (FDR=2.51e-07) but neither Adenoma (FDR>5e-02) nor Cerebellar Neoplasms (FDR>5e-02), which were associated to fewer genes as shown above. As expected, the list also included many other types of cancers related for example to lung, mouth, larynx, colon, hepatocytes, stomach, bones, prostate, or squamous cells.

Through this analysis, GS2D recapitulates that in the literature p53 is known for its role in many cancers but it is actually not often associated to adenoma (benign tumours) (e.g. [24]) or to multiple nervous system tumours (e.g. ependymomas [25] or non-astrocytic central nervous system tumours [26]).

Benchmarks and comparison

For 20 gold standard gene sets, we compared lists of enriched diseases produced by GS2D and ToppGene using CTD (ToppGene-CTD), OMIM (ToppGene-OMIM) or GWAS associations (ToppGene-GWAS) (Figure 2). The strict evaluation compared the ranks of the disease known to be related to the gene set and the relaxed evaluation compared the ranks of closely related diseases in the MeSH hierarchy (Table 1 and 2). For example, for the gold standard gene set known to be related to Inflammatory Bowel Diseases, the relaxed evaluation also considers ranks of Crohn Disease or Colitis Ulcerative (Table 3).

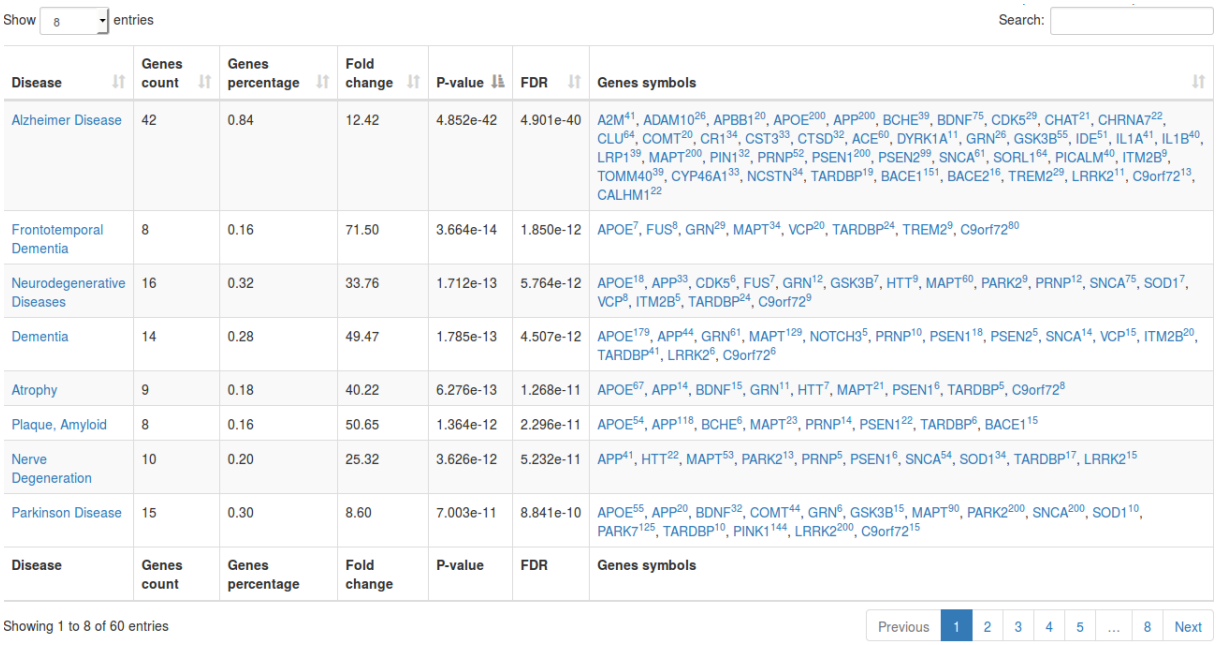
On strict evaluation of 10 GWAS gold standard gene sets, GS2D produced always the best ranking, nine times together with another method (Table 1). On relaxed evaluation of the 10 GWAS gold standard gene sets, GS2D produced always the best ranking, together with another method. On strict evaluation of 10 non-GWAS gold standard gene sets, GS2D produced eight times the best ranking, three times together with another method (Table 2). On relaxed evaluation of non-GWAS gold standard gene sets, GS2D produced always the best ranking, eight times together with another method.

Overall, GS2D was the best performing method in both the strict and relaxed evaluation. We evaluated CTD-based analyses as second best performing method. CTD contains considerably more gene-disease associations than the GWAS Catalog, OMIM and other curated datasets ([27] and DisGeNet database statistics [2]). Therefore, we compared the CTD data to automatically generated data used by GS2D.

GS2D defined more gene-disease associations than CTD (63503 and 23507, respectively). Whereas in the entire data GS2D covered a smaller number of diseases than CTD (2214 and 3880, respectively) (Table 4), the proportion of diseases associated to a single gene was strikingly smaller in GS2D (24.7% vs 60.3%). Overall, GS2D covered more diseases when considering only traits associated with at least 2 or more genes. As we consider meaningful only enrichment analyses based on diseases associated to more than one gene, we could conclude that for such analyses GS2D is more comprehensive.

DISCUSSION

We have implemented as a web tool a method called Gene set to Diseases (GS2D) performing disease enrichment analysis on human gene sets. Contrary to the most commonly used methods that use experimental or curated data to derive associations of genes to diseases, GS2D uses automatically derived gene-disease associations from co-occurrences in biomedical citations. When compared to results from similar tools that use different methods to associate genes to diseases, GS2D performed equally or better. The GS2D web tool can be used interactively from its simple and fast web interface or programmatically



Show 8 entries Search:

Disease	Genes count	Genes percentage	Fold change	P-value	FDR	Genes symbols
Alzheimer Disease	42	0.84	12.42	4.852e-42	4.901e-40	A2M[41], ADAM10[26], APBB1[20], APOE[200], APP[200], BCHE[39], BDNF[75], CDK5[29], CHAT[21], CHRNA7[22], CLU[64], COMT[20], CR1[34], CST3[33], CTSD[32], ACE[60], DYRK1A[11], GRN[26], GSK3B[55], IDE[51], IL1A[41], IL1B[40], LRP1[39], MAPT[200], PIN1[32], PRNP[52], PSEN1[200], PSEN2[99], SNCA[61], SORL1[64], PICALM[40], ITM2B[9], TOMM40[39], CYP46A1[33], NCSTN[34], TARDBP[19], BACE1[151], BACE2[16], TREM2[29], LRRK2[11], C9orf72[13], CALHM1[22]
Frontotemporal Dementia	8	0.16	71.50	3.664e-14	1.850e-12	APOE[7], FUS[8], GRN[29], MAPT[34], VCP[20], TARDBP[24], TREM2[9], C9orf72[80]
Neurodegenerative Diseases	16	0.32	33.76	1.712e-13	5.764e-12	APOE[18], APP[33], CDK5[6], FUS[7], GRN[12], GSK3B[7], HTT[9], MAPT[60], PARK2[9], PRNP[12], SNCA[75], SOD1[7], VCP[8], ITM2B[5], TARDBP[24], C9orf72[9]
Dementia	14	0.28	49.47	1.785e-13	4.507e-12	APOE[179], APP[44], GRN[61], MAPT[129], NOTCH3[5], PRNP[10], PSEN1[18], PSEN2[5], SNCA[14], VCP[15], ITM2B[20], TARDBP[41], LRRK2[6], C9orf72[6]
Atrophy	9	0.18	40.22	6.276e-13	1.268e-11	APOE[67], APP[14], BDNF[15], GRN[11], HTT[7], MAPT[21], PSEN1[6], TARDBP[5], C9orf72[8]
Plaque, Amyloid	8	0.16	50.65	1.364e-12	2.296e-11	APOE[54], APP[118], BCHE[6], MAPT[23], PRNP[14], PSEN1[22], TARDBP[6], BACE1[15]
Nerve Degeneration	10	0.20	25.32	3.626e-12	5.232e-11	APP[41], HTT[22], MAPT[53], PARK2[13], PRNP[5], PSEN1[6], SNCA[54], SOD1[34], TARDBP[17], LRRK2[15]
Parkinson Disease	15	0.30	8.60	7.003e-11	8.841e-10	APOE[55], APP[20], BDNF[32], COMT[44], GRN[6], GSK3B[15], MAPT[90], PARK2[200], SNCA[200], SOD1[10], PARK7[125], TARDBP[10], PINK1[144], LRRK2[200], C9orf72[15]
Disease	Genes count	Genes percentage	Fold change	P-value	FDR	Genes symbols

Showing 1 to 8 of 60 entries Previous 1 2 3 4 5 … 8 Next

Figure 3: Interactive output table. The figure shows the output table of a disease enrichment analysis on a gene set of 50 genes related to Alzheimer's disease. The enrichment analysis output can display a selected amount of diseases (top left-hand side drop-down menu), can be filtered (top right-hand side search box), sorted (clicks on column headers) and navigated by pages (bottom right-hand side navigation menu). Diseases are linked to the corresponding Medical Subject Headings thesaurus entries. Genes from the input gene set related to each disease are listed in the last column and linked to corresponding numbers (superscript numbers) of related PubMed citations.

	Rank on strict evaluation				Rank on relaxed evaluation			
GWAS gold standard gene set	GS2D	ToppGene-CTD	ToppGene-GWAS	ToppGene-OMIM	GS2D	ToppGene-CTD	ToppGene-GWAS	ToppGene-OMIM
Arthritis, Rheumatoid	1	1	1	1	1	1	1	1
Breast Neoplasms	1	1	1	-	1	1	1	-
Colitis, Ulcerative	1	1	1	-	1	1	1	-
Crohn Disease	1	1	1	-	1	1	1	-
Diabetes Mellitus, Type 2	1	1	1	1	1	1	1	1
Inflammatory Bowel Diseases	3	5	4	-	1	1	1	-
Lupus Erythematosus, Systemic	1	1	1	1	1	1	1	1
Multiple Sclerosis	1	1	1	2	1	1	1	2
Obesity	1	2	4	1	1	2	4	1
Prostatic Neoplasms	1	1	1	3	1	1	1	3

Table 1: Benchmarks on 10 GWAS gold standard gene sets. Ranks of the disease exactly matching the GWAS gold standard gene set (strict evaluation) or matching a closely related disease (relaxed evaluation) in results of enrichment analyses. Green cells denote best or equal best performance of Gene set to Diseases (GS2D) in comparison to the ToppGene web tool deriving gene-associations from the Comparative Toxicogenomics Database (ToppGene-CTD), the Genome-Wide Association Studies (ToppGene-GWAS), or the Online Mendelian Inheritance in Man (ToppGene-OMIM) database.

	Rank on strict evaluation				Rank on relaxed evaluation			
Non-GWAS gold standard gene set	GS2D	ToppGene-CTD	ToppGene-GWAS	ToppGene-OMIM	GS2D	ToppGene-CTD	ToppGene-GWAS	ToppGene-OMIM
Anemia, Aplastic	5	3	-	2	1	1	-	1
Cardiomyopathy, Hypertrophic, Familial	3	2	-	1	1	2	-	1
Ectodermal Dysplasia	1	2	-	2	1	1	-	1
Hair Diseases	2	-	-	-	1	1	-	1
Limb Deformities, Congenital	3	18	-	-	3	14	-	8
Malformations of Cortical Development	2	-	-	-	2	3	-	5
Mitochondrial Diseases	6	6	-	-	1	1	-	1
Peroxisomal Disorders	1	3	-	-	1	1	-	1
Spastic Paraplegia, Hereditary	1	1	-	1	1	1	-	1
Spinocerebellar Ataxias	1	1	-	1	1	1	-	1

Table 2: Benchmarks on 10 non-GWAS gold standard gene sets. Ranks of the disease exactly matching the non-GWAS gold standard gene set (strict evaluation) or matching a closely related disease (relaxed evaluation) in results of enrichment analyses using different types of gene-disease associations. Green cells denote best or equal best performance of Gene set to Diseases (GS2D) in comparison to the ToppGene web tool deriving gene-disease associations from the Comparative Toxicogenomics Database (ToppGene-CTD), the Genome-Wide Association Studies (ToppGene-GWAS), or the Online Mendelian Inheritance in Man (ToppGene-OMIM) database.

	GS2D	ToppGene-CTD	ToppGene-GWAS	ToppGene-OMIM
Rank 1	Crohn Disease	Colitis, Ulcerative	Crohn's disease	Rheumatoid Arthritis; RA
Rank 2	Colitis, Ulcerative	Crohn Disease	Coronary disease	Sarcoidosis, Susceptibility To, 1; SS1
Rank 3	**Inflammatory Bowel Diseases**	Arthritis, Rheumatoid	Type 1 diabetes	Human Immunodeficiency Virus Type 1, Susceptibility To
Rank 4	Arthritis, Rheumatoid	Lupus Erythematosus, Systemic	**Inflammatory bowel disease**	-
Rank 5	Diabetes Mellitus, Type 1	**Inflammatory Bowel Diseases**	Multiple sclerosis	-
Strict evaluation rank	3	5	4	-
Relaxed evaluation rank	1	1	1	-

Table 3: Benchmark of Inflammatory Bowel Diseases. The table lists top 5 results of disease enrichment analysis on a gold standard gene set known to be related to Inflammatory Bowel Diseases (in bold). The gold standard gene set was retrieved from the GWAS Catalog and contained 165 genes. Disease enrichment analysis was performed by Gene set to diseases (GS2D), or by the ToppGene service using gene-disease associations derived from the Comparative Toxicogenomics Database (ToppGene-CTD), the Genome-Wide Association Studies (ToppGene-GWAS), or the Online Mendelian Inheritance in Man (ToppGene-OMIM) database. For the strict evaluation, the rank of Inflammatory Bowel Diseases is retained. For the relaxed evaluation, either the rank of "Crohn Disease" or "Colitis, Ulcerative" (in red) is retained as they are directly related to Inflammatory Bowel Diseases in the MeSH hierarchical vocabulary.

Minimum number of genes per diseases	GS2D diseases	CTD diseases	Diseases in common
1	2214	3880	1558
2	1667	1539	1001
3	1389	1075	767
5	1044	690	510
10	676	404	304

Table 4: Comparison to the Comparative Toxicogenomics Database (CTD).

as a web service (results are usually returned in 1-2 seconds).

We have also demonstrated the use of GS2D in two use cases: (a) we have shown how knowledge about p53's role in cancers can be recapitulated from a set of genes interacting with p53 and (b) how tuning GS2D parameters helps get more stringent results for a set of genes related to Alzheimer's disease. Importantly, GS2D offers two ways to increase the stringency of the results: decreasing the FDR cutoff or increasing the minimal number of required co-occurrences. The former is usually also offered by comparable web tools (such as ToppGene), but not the latter that actually impacts the background set of gene-disease associations (reducing its size by keeping the most significant associations). Although such tuning is not applicable to CTD and OMIM (qualitative data), it would be interesting for enrichment analyses to filter out less confident GWAS associations using available quantitative values (e.g. p-values).

GS2D was compared to the ToppGene service because it used 3 different sources of gene-disease associations: CTD (ToppGene-CTD), OMIM (ToppGene-OMIM) and GWAS data (ToppGene-GWAS). Overall, GS2D performed better than ToppGene. More specifically, ToppGene-OMIM and ToppGene-GWAS performed worse than ToppGene-CTD. This may be explained by the bigger size of the CTD data ([27] and DisGeNet database statistics) or by its higher quality compared to GWAS data which is not always reproducible [28].

We have focused GS2D to human genes as we have found too limiting to discriminate the disease-related literature by species using co-occurrences (smaller resulting datasets; data not shown) and as disease-related studies are mainly about human diseases even if model species are involved.

In conclusion, GS2D is a fast and competitive web tool performing disease enrichment analysis on human gene sets. Since building its gene-disease associations takes a short time, regular updates can be planned. Its graphical and programmatic interfaces are accessible at: http://cbdm.uni-mainz.de/geneset2diseases.

ACKNOWLEDGEMENTS

We thank Dr. Taškova for critical reading of the manuscript.

AUTHOR CONTRIBUTIONS

JFF and MAAN designed the study and wrote the manuscript. JFF implemented the software and performed the experiments.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY DATA

High resolution figure files, together with supplementary items listed below, are available at Genomics and Computational Biology online.

Supplementary Table 1. Tab-separated values output file of gene-disease associations for 254 genes interacting with p53.

Supplementary Table 2. Tab-separated values output file of disease enrichment analysis of 254 genes interacting with p53.

ABBREVIATIONS

GS2D: Gene set to Diseases
CTD: Comparative Toxicogenomics Database
OMIM: Online Mendelian Inheritance in Man
GWAS: Genome-Wide Association Studies

REFERENCES

1. **Gene Ontology Consortium: going forward**. Nucleic Acids Research. 2014 nov;43(D1):D1049–D1056. doi:10.1093/nar/gku1179.
2. Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, Bohler A, et al. **WikiPathways: capturing the full diversity of pathway knowledge**. Nucleic Acids Res. 2015 oct;44(D1):D488–D494. doi:10.1093/nar/gkv1024.
3. **The ENCODE (ENCyclopedia Of DNA Elements) Project**. Science. 2004 oct;306(5696):636–640. doi:10.1126/science.1105136.
4. Soldatos TG, Perdigão N, Brown NP, Sabir KS, O'Donoghue SI. **How to learn about gene function: text-mining or ontologies?** Methods. 2015 mar;74:3–15. doi:10.1016/j.ymeth.2014.07.004.
5. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, et al. **The GOA database: Gene Ontology annotation updates for 2015**. Nucleic Acids Research. 2014 nov;43(D1):D1057–D1063. doi:10.1093/nar/gku1113.
6. Muro EM, Perez-Iratxeta C, Andrade-Navarro MA. BMC Bioinformatics. 2006;7(1):159. doi:10.1186/1471-2105-7-159.
7. Chen J, Bardes EE, Aronow BJ, Jegga AG. **ToppGene Suite for gene list enrichment analysis and candidate gene prioritization**. Nucleic Acids Research. 2009 may;37(Web Server):W305–W311. doi:10.1093/nar/gkp427.
8. Deng Y, Gao L, Wang B, Guo X. **HPOSim: An R Package for Phenotypic Similarity Measure and Enrichment Analysis Based on the Human Phenotype Ontology**. PLOS ONE. 2015 feb;10(2):e0115692. doi:10.1371/journal.pone.0115692.
9. Yu G, Wang LG, Yan GR, He QY. **DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis**. Bioinformatics. 2014 oct;31(4):608–609. doi:10.1093/bioinformatics/btu684.
10. Davis AP, Grondin CJ, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, et al. **The Comparative Toxicogenomics Database's 10th year anniversary: update 2015**. Nucleic Acids Research. 2015;43(D1):D914–D920. doi:10.1093/nar/gku935.
11. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. **The NHGRI GWAS Catalog, a curated resource of SNP-trait associations**. Nucleic Acids Research. 2013 dec;42(D1):D1001–D1006. doi:10.1093/nar/gkt1229.
12. Bravo À, Cases M, Queralt-Rosinach N, Sanz F, Furlong LI. **A Knowledge-Driven Approach to Extract Disease-Related Biomarkers from the Literature**. BioMed Research International. 2014;2014:1–11. doi:10.1155/2014/253128.
13. Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel HP. **Extraction of semantic biomedical relations from text using conditional random fields**. BMC Bioinformatics. 2008;9(1):207. doi:10.1186/1471-2105-9-207.
14. Lu Z, Kao HY, Wei CH, Huang M, Liu J, Kuo CJ, et al. **The gene normalization task in BioCreative III**. BMC Bioinformatics. 2011;12(Suppl 8):S2. doi:10.1186/1471-2105-12-s8-s2.
15. Wei CH, Kao HY, Lu Z. **GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains**. BioMed Research International. 2015;2015:1–7. doi:10.1155/2015/918710.
16. **Database resources of the National Center for Biotechnology Information**. Nucleic Acids Res. 2015 nov;44(D1):D7–D19. doi:10.1093/nar/gkv1290.
17. Yoav Benjamini YH. **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289–300.
18. R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria; 2013.
19. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. **Uncovering disease-disease relationships through the incomplete interactome**. Science. 2015;347(6224). doi:10.1126/science.1257601.
20. Fontaine JF, Priller F, Barbosa-Silva A, Andrade-Navarro MA. **Genie: literature-based gene prioritization at multi genomic scale**. Nucleic Acids Research. 2011 may;39(suppl):W455–W461. doi:10.1093/nar/gkr246.
21. Berlyand Y, Weintraub D, Xie SX, Mellis IA, Doshi J, Rick J, et al. **An Alzheimer's Disease-Derived Biomarker Signature Identifies Parkinson's Disease Patients with Dementia**. PLOS ONE. 2016 jan;11(1):e0147319. doi:10.1371/journal.pone.0147319.
22. Hussain M, Tian K, Mutti L, Krstic-Demonacos M, Schwartz JM. **The Expanded p53 Interactome as a Predictive Model for Cancer Therapy**. Genomics and Computational Biology. 2015 Sep;1(1):e20. doi:10.18547/gcb.2015.vol1.iss1.e20.
23. Hollstein M, Sidransky D, Vogelstein B, Harris C. **p53 mutations in human cancers**. Science. 1991 jul;253(5015):49–53. doi:10.1126/science.1905840.
24. Gandour-Edwards R, Kapadia S, Janecka I, Martinez A, Barnes L. **Biologic markers of invasive pituitary adenomas involving the sphenoid sinus**. Modern Pathology. 1995;8(2):160–164.
25. Fink KL, Rushing EJ, Schold SC, Nisen PD. **Infrequency of p53 gene mutations in ependymomas**. Journal of Neuro-Oncology. 1996;27(2):111–115. doi:10.1007/BF00177473.
26. Nozaki M, Tada M, Matsumoto R, Sawamura Y, Abe H, Iggo RD. **Rare occurrence of inactivating p53 gene mutations in primary non-astrocytic tumors of the central nervous system: reappraisal by yeast functional assay**. Acta Neuropathologica. 1998 mar;95(3):291–296. doi:10.1007/s004010050800.
27. Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI. **Gene-Disease Network Analysis Reveals Functional Modules in Mendelian, Complex and Environmental Diseases**. PLoS ONE. 2011 jun;6(6):e20284. doi:10.1371/journal.pone.0020284.
28. Nagai Y, Takahashi Y, Imanishi T. **VaDE: a manually curated database of reproducible associations between various traits and human genomic polymorphisms**. Nucleic Acids Research. 2014 oct;43(D1):D868–D872. doi:10.1093/nar/gku1037.

NOTES

[1]Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD, USA), 2016. World Wide Web URL: http://omim.org/

[2]DisGeNET, Integrative Biomedical Informatics Group, Research Programme on Biomedical Informatics (Barcelona, Spain), 24 August 2016. World Wide Web URL: http://www.disgenet.org/web/DisGeNET/menu/dbinfo#sources