OXFORD UNIVERSITY PRESS | DATABASE

DISEASES 2.0: a weekly updated database of disease--gene associations from text mining and data integration

Journal:	*DATABASE*
Manuscript ID	DATABASE-2021-0143.R1
Manuscript Type:	Database Update
Date Submitted by the Author:	14-Feb-2022
Complete List of Authors:	Grissa, Dhouha; University of Copenhagen Novo Nordisk Foundation Center for Protein Research, Disease Systems Biology Junge, Alexander; University of Copenhagen Novo Nordisk Foundation Center for Protein Research, Disease Systems Biology Oprea, Tudor; University of New Mexico School of Medicine, Department of Internal Medicine Jensen, Lars; Department of Disease Systems Biology, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen
Keywords:	Named entity recognition, Data integration, Web resource, Text mining, Gene-disease association, Database

DISEASES 2.0: a weekly updated database of disease–gene associations from text mining and data integration

Dhouha Grissa[1], Alexander Junge *[1], Tudor I. Oprea[1,2], and Lars Juhl Jensen[1]

[1]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200, Copenhagen, Denmark.
[2]Department of Internal Medicine, Division of Translational Informatics, University of New Mexico Health Sciences Center, Albuquerque, NM, USA.

Abstract

The scientific knowledge about which genes are involved in which diseases grows rapidly, which makes it difficult to keep up with new publications and genetics datasets. The DISEASES database aims to provide a comprehensive overview by systematically integrating and assigning confidence scores to evidence for disease–gene associations from curated databases, genome-wide association studies (GWAS), and automatic text mining of the biomedical literature. Here, we present a major update to this resource, which greatly increases the number of associations from all these sources. This is especially true for the text-mined associations, which have increased by at least 9-fold at all confidence cutoffs. We show that this dramatic increase is primarily due to adding full-text articles to the text corpus, secondarily due to improvements to both the disease and gene dictionaries used for named entity recognition, and only to a very small extent due to the growth in number of PubMed abstracts. DISEASES now also makes use of a new GWAS database, TIGA, which considerably increased the number of GWAS-derived disease–gene associations. DISEASES itself is also integrated into several other databases and resources, including GeneCards/MalaCards, Pharos/TCRD, and the Cytoscape stringApp. All data in DISEASES is updated on a weekly basis and is available via a web interface at `https://diseases.jensenlab.org`, from where it can also be downloaded under open licenses.

*Present address: Corti, Copenhagen, Denmark

1 Introduction

To understand human diseases at the molecular level, we need a comprehensive overview of which genes are linked to each disease. Since these links can come from many types of data, each of which is growing at a fast rate, there is a need for frequently updated databases that integrate the heterogeneous evidence for disease–gene associations. To this end, we provide the freely available DISEASES database resource [1], which has been continuously updated on a weekly basis since 2014 on an in-house server. This resource automatically extracts disease–gene associations from the biomedical literature by identifying mentions of disease and gene names and counting how often they are co-mentioned. These are combined with manually curated associations and experimental evidence from genome-wide association studies (GWAS).

The first version of DISEASES included text mining of 24 million abstracts available from the PubMed database, which has since grown by another 8 million new abstracts. Moreover, text mining of full-text articles can yield approximately 50% more disease–gene associations (at the same false positive rate) than text mining of the corresponding abstracts [2]. Combined with the rapid growth of open-access publishing [3], which allows for text mining and redistribution, this shows a clear opportunity for improving resources like DISEASES to go beyond text mining of abstracts only.

Text mining has been applied to many tasks in the biomedical domain, such as identifying genes and other named entities [4] in text and subsequently extracting associations between genes and other genes [5], pathways [6], and diseases [7]. Many studies have focused on doing the latter based on biomedical abstracts only [8], whereas fewer have included full-text articles too [9]. General text-mining efforts, not specifically focused on disease–gene associations, of course also extract such associations [10, 11].

DISEASES is not the only database to gather evidence of disease–gene associations. Such associations have for many years been manually annotated by curators of both general protein databases, such as UniProtKB/Swiss-Prot [12], and databases focused on genetics of diseases, such as Online Mendelian Inheritance in Man (OMIM) [13] and MedlinePlus from Genetics Home Reference (GHR) [14]. In addition to these broad databases, many specialist databases exist which focus on specific diseases or classes of diseases, including the cancer mutation databases Catalog of Somatic Mutations in Cancer (COSMIC) [15] and intOGen [16].

Genome-wide association studies (GWAS) are another important source of disease–gene associations, which comes with its own ecosystem of database resources. In GWAS, statistically significant associations between single nucleotide polymorphisms (SNPs) and phenotypic traits (including diseases) are identified and used to infer gene–trait associations. These associations, both at the SNP and gene level, are collected by the National Human

Genome Research Institute (NHGRI)–European Bioinformatics Institute (EBI) GWAS Catalog [17], GWAS Central [18] databases, and GWASdb [19]. However, GWAS results are complex to interpret; the identified SNPs are typically not the causal mutations, which due to linkage disequilibrium could reside anywhere within a chromosomal region that may contain multiple genes [20]. Several secondary GWAS databases, such as DistiLD [21] and TIGA [22], thus aim to help non-experts interpret GWAS results by integrating other relevant information and prioritizing the results.

Several integrative database resources, like DISEASES, combine many of the types of evidence for disease–gene associations mentioned above. The best known of these are probably MalaCards/GeneCards [23] and DisGeNET [24], which take two very different approaches. The MalaCards and GeneCards databases present one page for each disease and gene, respectively, which gives the user a very comprehensive overview of the available information, including text mining from DISEASES. DisGeNET, on the other hand, has a strong focus on scoring the associations and on making data available amenable to further computational analysis via application programming interfaces (APIs). The latter is also true for knowledge-based aggregators such as Pharos [25] and Open Targets [26], which integrate many types of evidence from numerous sources, including gene–disease associations.

In this paper, we describe the major improvements of the DISEASES resource made since the 2015 publication [1]. The gene set and associated dictionary have been updated to be consistent with the latest version of STRING [27] and the text corpus in DISEASES is now automatically constructed by merging the open-access subset of PubMed Central (PMC) with PubMed abstracts. This has jointly led to substantial improvements of the text-mining results. We have further updated DISEASES to import experimental data from a new GWAS resource, TIGA [22]. We map all disease–gene associations to a common set of identifiers and provide confidence scores for the associations, which are comparable across evidence types. All data are freely available both via a web interface (`https://diseases.jensenlab.org/`), as bulk download files, and through integration into other resources and tools, specifically Cytoscape, TIN-X, and Pharos.

2 Materials & Methods

The DISEASES database combines heterogeneous evidence from several sources. We will go through these, starting with three databases of manually curated disease—gene associations, followed by two sources of experimental evidence, and finally the automatic text mining, which we break down into corpus construction, dictionary construction, named entity recognition

(NER), and co-occurrence scoring. Finally, we describe how the DISEASES confidence scores are assigned for each type of evidence.

2.1 MedlinePlus

The genetics section of the MedlinePlus resource, formerly known as Genetics Home Reference, includes disease–gene associations obtained from manual curation of the biomedical literature [14]. We first download the list of all diseases and then query the MedlinePlus REST API with each disease to retrieve the list of associated HGNC gene symbols. We then use the dictionaries described later to map the disease names and gene symbols to their Disease Ontology identifiers [28] and STRING v11 identifiers, respectively.

2.2 UniProt Knowledgebase (UniProtKB)

The Swiss-Prot section of UniProtKB consists of expert-reviewed protein entries, which include diseases associated with each protein among many other types of biological information [12]. We extract the diseases associated with a protein entry by parsing the keyword field, where they are specified using a controlled vocabulary. We manually mapped these to their corresponding concepts in Disease Ontology and used the dictionary described later to map the UniProtKB accession numbers to STRING v11 identifiers.

2.3 Amyloidoses Collection (AmyCo)

The AmyCo database specifically collects information on amyloidoses and other diseases related to amyloid deposition [29]. It contains manually curated data from 249 articles on 75 diseases classified into two broad groups: amyloidoses and clinical conditions associated with amyloidosis, including precursors and proteins co-deposited with amyloid deposits. AmyCo identifiers are mapped to their corresponding Disease Ontology identifiers whenever the AmyCo name could be found as an exact synonym; otherwise the AmyCo identifier is mapped to the Disease Ontology broader parent(s).

2.4 Target Illumination by GWAS Analytics (TIGA)

TIGA [22] is a new weekly updated web resource that imports GWAS data from the NHGRI-EBI GWAS Catalog [17], maps SNPs to the nearest protein-coding genes, and evaluates the confidence of each gene–trait association. The latter is done by calculating an average rank score based on the number and distance of SNPs supporting the association, the p-value of the most significant SNP, and the number of studies supporting the association weighted by the Relative Citation Ratio [30] of the underlying publications. From TIGA, we extract the subset of traits that are diseases and map their Experimental Factor Ontology (EFO) terms to the corresponding Disease

Ontology terms based on ontology cross-references and the EMBL-EBI Ontology Xref Service. The Ensembl gene identifiers are mapped to STRING v11 identifiers using the gene dictionary described later.

2.5 Text corpora

As the starting point for doing text mining, a large body of biomedical texts is needed. We compile such a corpus based on the PMC open-access subset [3], which consists of 7.3 million full-text articles, and the PubMed abstract database, which contains 39 million entries, 22 million of which have an English abstract. To construct a combined corpus, we download both PMC and PubMed in XML format, specifically, the BioC version of PMC [31] and the PubMed baseline plus daily updates. The latter updates also mean that retracted articles are automatically removed from the corpus as soon as they have been marked as such in PubMed. As a last step, we exclude 826 publications, which are believed to contain falsified data and to have been created by several recently discovered paper mills [32].

As PMC contains articles in several languages and the text-mining pipeline in DISEASES is designed only for English text, we use the pretrained language detection models from fastText [33] to identify the language of each PMC article and remove articles not in English. We next run the NER software (described later) on the English-language articles to count the number of unique entities found in each article. To eliminate articles that mention long lists of genes or diseases, we removed the 630 PMC articles that mentioned more than 200 genes or diseases. For quality reasons and to have consistent article metadata, we decided to only include articles from PMC that are indexed in PubMed. We thus used the identifier mapping file from PMC to convert PMCIDs to PMIDs, and discarded all PMC articles for which a PMID did not exist. For the remaining articles, we merged the information from PMC and PubMed, using the metadata, title and abstract from PubMed and extending it with the article body text from PMC. Where no PMC open-access version of an article was available, we simply used the metadata, title, and abstract from PubMed.

The corpus in DISEASES is updated every weekend. All results presented in this paper are based on October 20th, 2021 version of the corpus (FullText2021). To assess the impact of including full-text open-access publications, we also have a second corpus (PubMed2021), which includes only the title and abstract text of the same publications. Finally, to be able to assess how much the general growth of PubMed contributes to the performance, we have a third corpus, which consists of $\sim$ 24 million abstracts published by end of 2013 (PubMed2013). This corpus closely resembles what was in the initial version of DISEASES, which was submitted in January 2014 [1].

2.6 Dictionaries

For mapping names and identifiers and for recognizing them in text, we need comprehensive dictionaries of human genes and diseases. The dictionary of diseases is constructed based on all the names and synonyms from Disease Ontology [28] and extended with additional amyloidoses from AmyCo, mappings to ICD-10, and manual additions of missing disease synonyms and acronyms. The human gene dictionary was obtained from STRING v11.0 [27] and is based on information from Ensembl [34], UniProtKB [12], and HGNC [35] databases. We further automatically eliminate clashes between HGNC gene symbols and disease names and extend both dictionaries with orthographic variations of names using the exact same rules as in the first version of DISEASES [1].

We frequently update the dictionaries to incorporate changes to Disease Ontology and to correct errors identified by users. In this paper we make use of two frozen versions of the dictionaries, namely one from October 20th, 2021 (Dict2.0) and, for comparison, the dictionaries from the first version of DISEASES [1] (Dict1.0). The latest dictionary is available from the Downloads tab of the DISEASES web resource.

2.7 Named entity recognition (NER)

To do NER on the very large text corpora described above — and make frequent updates feasible — a highly efficient tool for matching the dictionaries against the text is needed. As in the previous version of DISEASES, we do this using the Tagger software, which is described in detail elsewhere [36] (`https://github.com/larsjuhljensen/tagger`), is available for download from the DISEASES website. Briefly, the combined dictionary is first loaded into memory in a custom hash table that allows fast, case-insensitive lookup and further allows for arbitrary insertion and deletion of hyphens. We then tokenize the text on white-space and special characters (including hyphen and slash) and look up combinations of tokens in the combined dictionary to identify left-most longest matches. To improve the precision, we globally block tagging of names that would otherwise give rise to many false positives by manually inspecting the tagging results of all names that occur more than 2000 times in PubMed as well as names that gave rise to errors reported by users. Each match in the text is normalized to the unique entity identifier from the dictionary and, in case of diseases, the term is backtracked to all parent terms in Disease Ontology.

2.8 Co-occurrence scoring

From the NER results, we calculate co-occurrence score between any given pair of a gene and a disease, which quantifies how much these entities have been mentioned together in the text corpus. The scoring scheme takes into account that co-occurrences within sentences are stronger evidence than co-occurrences across sentences within a paragraph, which in turn are stronger than co-occurrences across paragraphs within a paper. The scoring scheme further takes into account both how much the entities co-occur on an absolute scale and relative to what would be expected by random chance (z-scores). This approach is that same as was used in DISEASES v1 [1], except that the scoring having been extended to handle full-text articles as previously described [2].

2.9 Comparison to DISEASES v1

To facilitate comparison with DISEASES v1, which differs both in terms of the dictionaries and the text corpus used, we generated four datasets of text-mining results. These represent the following combinations of the old and new dictionaries (*Dict1.0* and *Dict2.0*) with text corpora representing the abstracts used in DISEASES v1 (*PubMed2013*), the full set of abstracts now available (*PubMed2021*), and the combined corpus including also full-text articles from PMC (*FullText2021*):

(i) *Dict1,PubMed2013*: *PubMed2013* mined using *Dict1.0*, representing the text-mining channel of DISEASES v1 when published

(ii) *Dict1,PubMed2021*: *PubMed2021* mined using *Dict1.0* to show the effect of updating DISEASES v1 with new abstracts

(iii) *Dict2,PubMed2021*: *PubMed2021* mined using *Dict2.0* to capture the changes attributed to dictionary improvements

(iv) *Dict2,FullText2021*: *FullText2021* mined using *Dict2.0*, representing the text-mining channel of DISEASES v2

2.9.1 Gold standard of disease–gene associations

To build an up-to-date gold standard of disease–gene associations, we followed the approach described in the original DISEASES publication [1]. We used the knowledge channel of manually curated annotations imported from UniProtKB and MedlinePlus (doi:10.6084/m9.figshare.17075708). A new reduced benchmarking set is then generated that includes only entity names explicitly annotated in UniProtKB and MedlinePlus and, with diseases broader parent terms. When an annotated disease was a child term of another annotated disease, we kept the broader parent terms and backtracked the child-term annotations to it via the is_a relationships in the ontology. The final gold standard comprises $7,005$ of inferred disease–gene associations, all were given a high confidence score of 4 to 5 stars to show how they are biologically meaningful.

2.9.2 Benchmarking of text-mined associations

To evaluate and compare the quality of the text-mining results, we benchmarked each of the four sets of text-mining results (*Dict1.0_PubMed2013*, *Dict1.0_PubMed2021*, *Dict2.0_PubMed2021*, and *Dict2.0_FullText2021*) on the gold standard of disease–gene associations. Given a disease–gene association, we labeled it as positive if the association exists in the gold standard, labeled it as negative if both the disease and the gene (but not the association) exist in the gold standard, and otherwise discarded it. Based on this binary labeling, we constructed receiver operating characteristic (ROC) curves for each of the four sets of text-mining results by sorting the associations descending by score and plotting the true positive rate (TPR) against the false positive rate (FPR). To quantify the difference between the ROC curves, we calculated the area under curve (AUC) for each.

2.9.3 Benchmarking of GWAS-derived associations

The quality of TIGA gene–disease associations has been discussed in detail in [22]. Briefly, a benchmarking was performed against the same gold standard described above, yielding a ROC curve with an AUC of 0.73.

2.9.4 Confidence scores

To make the different types of evidence comparable, we assign confidence scores on a scale from 1 to 5 stars. The manually curated gene–disease associations in the knowledge channel, which also make up the gold standard, were assigned the highest confidence of 5 stars.

As in the first version of DISEASES [1], the co-occurrence text-mining results in a z-score for each association, which is converted into the final confidence scores using a heuristic. Each star corresponds to two standard deviations above random expectation, that is a z-score of 6 would become 3 stars. As automatic text mining is never as reliable as manual curation, the confidence scores for text mining are capped at 4 stars.

The gene–disease association scores from TIGA are made comparable to these other scores based on the benchmarks described above. The benchmark of TIGA gives us the relationship between TIGA scores and probability of association in the gold standard. The benchmark of text mining similarly gives us the association between z-scores and probability of association in the gold standard. This allows us to convert TIGA scores into probabilities, which are in turn converted to the corresponding z-scores and finally to the star confidence scores.

3 Results & Discussion

3.1 Overview of the DISEASES resource

Figure 1 gives an overview of the disease–gene associations in DISEASES by mapping them onto high-level terms in Disease Ontology that have at least 100 associated genes. For each disease term, the area shows the number of associations with a confidence score of at least 3 stars, which is further broken down based on the source of the associations (knowledge, experiments, or text mining). Automatic text mining is by far the biggest source of associations for all diseases, accounting for more than 60% of the total.

The category of diseases with the most associations, especially from knowledge and experiments, is *disease of anatomical entity*, within which we see a fairly even distribution across many anatomical systems. This is followed by *disease of cellular proliferation*, which almost exclusively covers cancer–gene associations. By contrast, we find quite few associations (7443) for *disease of infectious agent*, which come exclusively from text mining. The latter is to be expected, since the DISEASES database aims to identify causal associations between genes and diseases as reflected in the choice of gold standard. We thus do not aim to link a gene to a disease based on just change in expression.

3.2 Growth of the number of disease-gene associations

The number of open access articles available from PMC has been grown exponentially over time, reaching more than 7 million as of September 2021. The inclusion of these — as well as the more than 8 million new PubMed abstracts published since the initial release of DISEASES in 2015 — has an obvious and direct effect on the number of disease–gene association one can find by text mining. While the text-mining channel in the DISEASES database is our main focus in this article, it is not the only improvement. Table 1 provides an overview of the content of the original and the new releases DISEASES database, showing the number of genes, diseases, and associations provided by each evidence channel.

The *Knowledge* channel has more than doubled in terms of both disease–gene associations and unique diseases covered. This growth comes primarily from GHR, which has in the meantime been integrated into MedlinePlus, but UniProtKB has also grown substantially. AmyCo contributes a comparably low number of new associations, since it covers only a specific type of diseases.

The *Experiments* channel has changed in many ways between the two versions. Replacing DistiLD with TIGA has increased the number GWAS-based associations by more than a factor of four and more than doubled the coverage of genes. However, with the new release we have also to remove COSMIC for license reasons, thus losing more than half of the experimental

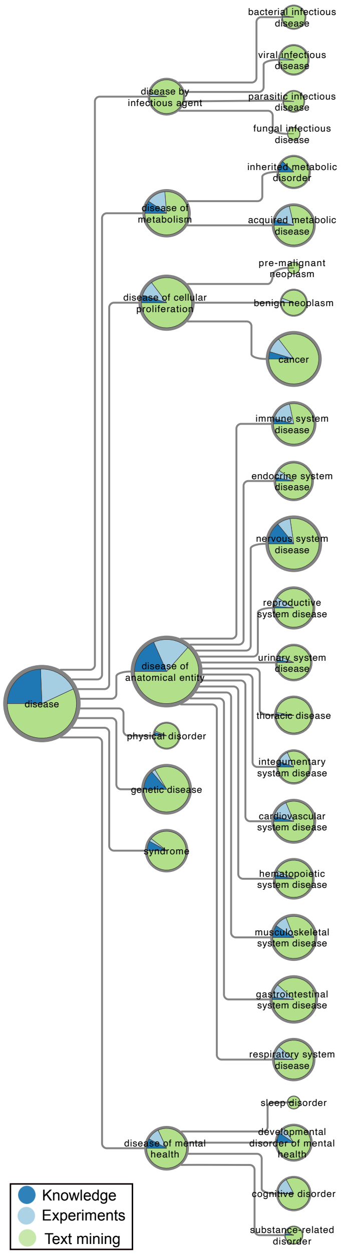

Figure 1: **Overview of disease–gene associations in DISEASES.** The number of disease–gene associations with a confidence score of at least 3 stars is proportional to the area of the pie charts, which represent high-level terms from Disease Ontology. In each pie chart, the associations are broken down by evidence type, i.e. curated knowledge, GWAS experiments, and automatic text mining of the literature.

associations in the original release of DISEASES. All in all, the *Experiments* channels has grown by over 70%.

For the *Text mining* channel, we subdivide the evidence by their confidence scores as presented in the DISEASES web interface. That is, we present the number of genes, diseases, and associations that rated as 2-, 3-, and 4-star confidence as well as the total numbers for the channel (including association scoring below 2 stars). The number of associations has increased dramatically at all confidence levels and especially at high confidence levels, with over 9-fold increase overall and over 17-fold increase for 4-star associations. The same trend holds true when looking at the numbers of unique genes and diseases covered.

Evidence channel	**Associations**		**Genes**	**Diseases**	
Knowledge					
DISEASES 2021	36,448	(6,739)	3,723	1,528	(1,048)
AmyCo	1,225	(186)	57	180	(52)
MedlinePlus	23,957	(3,516)	2,383	1,460	(966)
UniProtKB	19,517	(3,739)	2,642	270	(119)
DISEASES 2015	15,231	(2,953)	2,001	735	(453)
GHR	7,551	(1,169)	965	671	(390)
UniProtKB	11,576	(2,187)	1,651	271	(120)
Experiments					
DISEASES 2021	152,611	(26,346)	9,180	574	(295)
TIGA	152,611	(26,346)	9,180	574	(295)
DISEASES 2015	89,073	(20,206)	10,711	423	(264)
COSMIC	55,791	(13,050)	8,786	142	(76)
DistiLD	36,650	(7,185)	4,315	351	(210)
Text mining					
DISEASES 2021	4,512,870		19,116	8,537	
4 star confidence	18,129		2,988	2,959	
3 star confidence	224,642		11,207	6,711	
2 star confidence	1,659,331		18,913	8,342	
DISEASES 2015	478,407		15,631	4,598	
4 star confidence	1,044		478	662	
3 star confidence	15,226		3,207	2,267	
2 star confidence	142,892		12,706	4,354	

Table 1: **Comparison of the new and original versions of DISEASES.** For each evidence channel we show the number of associations, unique genes, and unique diseases for both the new and the originally published version of DISEASES. In case of the *knowledge* and *experiments* channels, these numbers are further provided for each of the source databases. The numbers in parentheses are the counts before evidence was backtracked to parent terms in Disease Ontology. For the *text mining* channel, we instead subdivide the counts by confidence score.

3.3 Improved quality of text-mined associations

We assessed the quality of the disease–gene associations from the new version of DISEASES to the originally published version by benchmarking both against a gold standard of manually annotated gene–disease associations (see Materials & Methods for details). The results are shown as ROC curves in Figure 2, which reveals a substantial improvement both overall (AUC increasing from 0.829 to 0.916) and in the low false-positive-rate part, which is arguably the most relevant part. As the ROC curve for the new version is consistently above that of the original one, the new version constitutes an improvement regardless of whether the user cares most about getting higher true positive rate or lower false positive rate.

This performance improvement is due to a combination of i) general growth in the number of biomedical abstracts available from PubMed, ii) improvements to the dictionaries used for NER, and iii) the addition of full-text articles from the PMC open access subset. To quantify the importance of each of these factors, we show two additional ROC curves in Figure 2: performance when updating with new abstracts but still using the original dictionaries and performance when further updating the dictionaries. Comparing the four ROC curves shows that the growth of PubMed abstracts alone gives only a small improvement of the AUROC from 0.845 to 0.859. The use of the new dictionaries leads to a bigger incremental improvement, increasing the AUROC from 0.859 to 0.866. However, the addition of full-text articles to the corpus is responsible for the biggest improvement, bringing the AUROC from 0.866 to = 0.922.

These results show that while the growth of the literature does give an almost free improvement, only requiring the pipeline to rerun on latest PubMed, the vast majority of the improvement seen between the original version of DISEASES and the new version stems from our work on improving the dictionaries used for NER and on integrating full-text articles into the corpus. The results also highlight how important it is for text mining efforts to be permitted to process full-text articles rather than only abstracts.

3.4 Research paper mills

To the best of our knowledge, all assessment of text-mining results to date have focused purely on the ability of a text-mining system to correctly extract what is stated in the text. However, from the perspective of using text mining to construct a knowledgebase from literature, is equally important if what is stated in the text is true. Co-mentioning-based systems indirectly address this, since high-scoring associations will be supported by multiple publications.

Recently, the problem of incorrect information in the literature has become a bigger concern due to the discovery of so-called ”paper mills”. These appear to be companies that mass produce fake articles and sell them to researchers at Chinese hospitals [37]. As these articles were published in international journals indexed in PubMed, they would by default be included in our text corpus, thus providing false support for disease–gene associations in our database. To avoid this, we have compiled a list of the 826 papers iden-

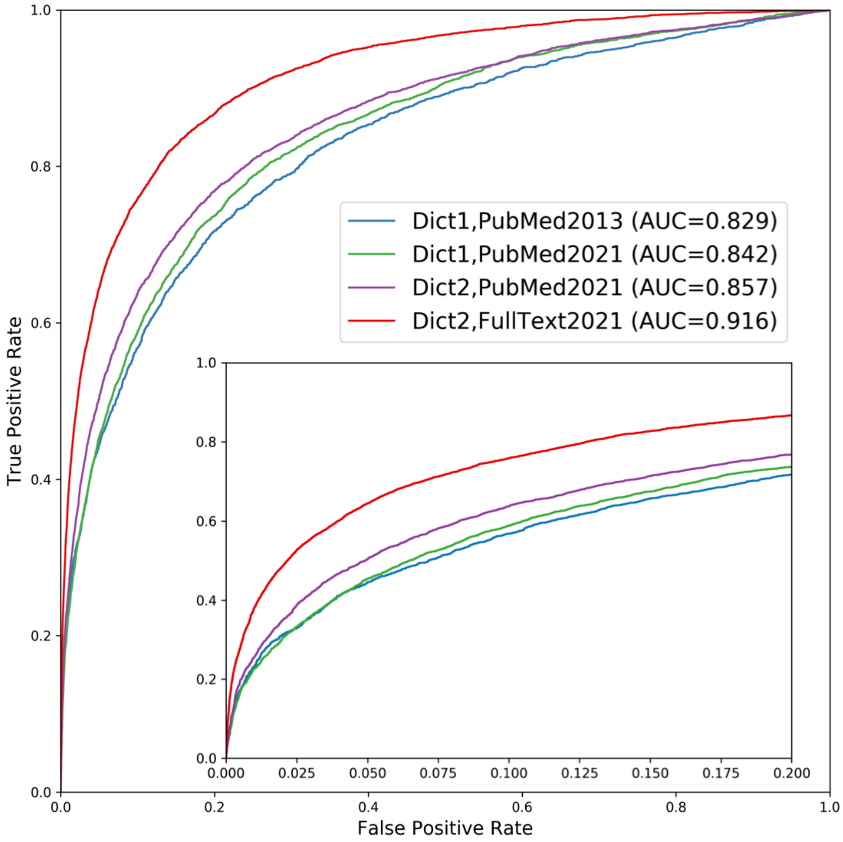

Figure 2: **Performance improvement of the *text mining* channel.** As shown in the receiver operating characteristic (ROC) curves, text mining performs markedly better in the new version of DISEASES (Dict2,FullText2021) compared to the originally published one (Dict1,PubMed2013). To quantify the sources of improvements, we show two additional curves: one using the new dictionary on the latest abstract collection only (Dict2,PubMed2021), and another using the old dictionary on the same abstracts (Dict1,PubMed2021). Comparing the curves reveals that most of the improvement stems from the addition of full-text articles, but that the new disease and gene dictionaries also led to considerable improvement. By contrast, the growth in PubMed abstracts from 2013 to 2021 made only a minor difference. The insert shows a zoom of the high-confidence part of the plot.

tified so far to originate from paper mills (as per June 2021) and explicitly exclude these from our text corpus. To allow others to also easily exclude these papers, the latest list is available for download from the DISEASES website.

3.5 Integration into other resources

Just like DISEASES itself builds upon other databases, we have designed it to be easy to integrate into other resources. We do this both from a technical perspective by providing simple bulk download options and from a legal perspective by not integrating any data that would prevent us using an open license. Several tools and databases already take advantage of this, importing either disease–gene associations from all evidence channels or specifically the associations from text mining.

The GeneCards and MalaCards databases, both members of the GeneCards suite, provide a comprehensive overview of information on human genes, including diseases associations, by integrating evidence from 150 sources [38]. One of these is the text-mined disease–gene associations from DISEASES, which GeneCards/MalaCards downloads on a regular basis and combines with associations from other source databases. The GeneCards/MalaCards web resources link back to DISEASES website to allow users to easily inspect the text-mining evidence for any given association. The Target Central Resource Database (TCRD) and the associated Pharos [25] and TIN-X web resources [39], which aim to shed light on potential new drug targets, similarly obtain up-to-date disease–gene associations from the DISEASES database. DISEASES (and "tagger" output) are also an integral part of Geneshot [40] and Harmonizome [41].

DISEASES is also designed to interface easily with the STRING, COMPARTMENTS and TISSUES resources by using the same gene identifiers. Through the Cytoscape app *stringApp*, it is thus possible to quickly retrieve a human protein network for any disease of interest [42]. To do this, stringApp first queries the DISEASES database to obtain a list of genes associated with the disease and subsequently queries STRING to obtain the corresponding protein network.

4 Conclusion

The DISEASES database has since 2014 provided the community with a weekly updated resource of disease–gene associations. The latest version features several important improvements compared to the original publication. In addition to text mining PubMed, which has meantime grown by another 8 million abstracts, the DISEASES text corpus now also includes open access full-text articles from PubMed Central. Together with technical improvements to the text-mining pipeline itself, this has led to a $>$ 9-fold increase in the number of disease–gene associations extracted at any confidence cutoff. DISEASES has also been upgraded to use GWAS data via the new TIGA database [22], which increased the number of experimental associations by more than 70%.

The database is freely available at `https://diseases.jensenlab.org/`

where it can be browsed via a web interface as well as downloaded in its entirety to facilitate large-scale analysis. Moreover, DISEASES is designed to integrate easily with other resources, and the disease–gene associations are available through other resources, including the GeneCards/MalaCards and TCRD/Pharos databases, Harmonizome, Geneshot, and the Cytoscape stringApp.

Acknowledgements This work was supported by US National Institutes of Health grant U24 224370 for ”Illuminating the Druggable Genome Knowledge Management Center” (IDG KMC), and by the Novo Nordisk Foundation (grant number NNF14CC0001).

References

[1] S. Pletscher-Frankild, A. Pallejã, K. Tsafou, J. X. Binder, L. J. Jensen, DISEASES: Text mining and data integration of disease-gene associations, Methods 74 (2015) 83–89. doi:10.1016/j.ymeth.2014.11.020.

[2] D. Westergaard, H.-H. Stærfeldt, C. Tønsberg, L. J. Jensen, S. Brunak, A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts, PLOS Computational Biology 14 (2) (2018) 1–16. doi:10.1371/journal.pcbi.1005962.

[3] D. C. Comeau, C.-H. Wei, R. Islamaj Doğan, Z. Lu, PMC text mining subset in BioC: about three million full-text articles and growing, Bioinformatics 35 (18) (2019) 3533–3535. doi:10.1093/bioinformatics/btz070.

[4] M.-T. Pandi, P. J. van der Spek, M. Koromina, G. P. Patrinos, A novel text-mining approach for retrieving pharmacogenomics associations from the literature, Frontiers in pharmacology 11 (2020) 602030. doi:10.3389/fphar.2020.602030.

[5] I. Karadeniz, J. Hur, Y. He, A. Özgür, Literature mining and ontology based analysis of host-Brucella gene-gene interaction network, Frontiers in microbiology 6 (2015) 1386. doi:10.33899/fmicb.2015.01386.

[6] X. Qin, X. Yao, J. Xia, A novel metric to quantify the effect of pathway enrichment evaluation with respect to biomedical text-mined terms: Development and feasibility study, JMIR medical informatics 9 (2021) e28247. doi:10.2196/28247.

[7] M. Simmons, A. Singhal, Z. Lu, Text Mining for Precision Medicine: Bringing Structure to EHRs and Biomedical Literature to Understand

Genes and Health., Advances in experimental medicine and biology 939 (2016) 139–166. doi:0.1007/978-981-10-1503-8_7.

[8] J. Zhou, B.-Q. Fu, The research on gene-disease association based on text-mining of PubMed, BMC Bioinformatics 19 (2018) 37. doi:10.1186/s12859-018-2048-y.

[9] J. Czarnecki, A. J. Shepherd, Mining Biological Networks from Full-Text Articles, Springer New York, New York, NY, 2014, pp. 135–145. doi:10.1007/978-1-4939-0709-0_8.

[10] T. K. Jenssen, A. Laegreid, J. Komorowski, E. Hovig, A literature network of human genes for high-throughput analysis of gene expression, Nature genetics 28 (1) (2001) 21–8. doi:10.1038/ng0501-21.

[11] Y. Tsuruoka, J. Tsujii, S. Ananiadou, FACTA: a text search engine for finding associated biomedical concepts, Bioinformatics 24 (21) (2008) 2559–60. doi:10.1093/bioinformatics/btn469.

[12] The UniProt Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Research 46 (D1) (2018) D158–D169. doi:10.1093/nar/gkw1099.

[13] J. S. Amberger, C. A. Bocchini, A. F. Scott, A. Hamosh, OMIM.org: leveraging knowledge across phenotype–gene relationships, Nucleic Acids Research 47 (D1) (2018) D1038–D1043. doi:10.1093/nar/gky1151.

[14] C. Fomous, J. A. Mitchell, A. McCray, Genetics Home Reference: Helping patients understand the role of genetics in health and disease, Community Genet 9 (2006) 274–278. doi:10.1159/000094477.

[15] S. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J. Teague, P. Futreal, M. Stratton, The Catalogue of Somatic Mutations in Cancer (COSMIC), Current protocols in human geneticsdoi:10.1002/0471142905.hg1011s57.

[16] F. Martínez-Jiménez, F. Muiños, I. Sentís, J. Deu-Pons, I. Reyes-Salazar, C. Arnedo-Pac, L. Mularoni, O. Pich, J. Bonet, H. Kranas, A. Gonzalez-Perez, N. Lopez-Bigas, A compendium of mutational cancer driver genes, Nature reviews. Cancer 20 (10) (2020) 555–572. doi:10.1038/s41568-020-0290-x.

[17] A. Buniello, J. A. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, McMahon, et al., The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019, Nucleic Acids Research 47 (D1) (2018) D1005–D1012. doi:10.1093/nar/gky1120.

[18] T. Beck, T. Shorter, A. J. Brookes, GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies, Nucleic Acids Research 48 (D1) (2019) D933–D940. doi:10.1093/nar/gkz895.

[19] M. J. Li, Z. Liu, P. Wang, M. P. Wong, M. R. Nelson, J.-P. A. Kocher, M. Yeager, P. C. Sham, S. J. Chanock, Z. Xia, J. Wang, GWASdb v2: an update database for human genetic variants identified by genome-wide association studies, Nucleic Acids Research 44 (D1) (2015) D869–D876. doi:10.1093/nar/gkv1317.

[20] K. A. Frazer, S. S. Murray, N. J. Schork, E. J. Topol, Human genetic variation and its contribution to complex traits, Nature reviews. Genetics 10 (4) (2009) 241–251.

[21] A. Pallej/'a, H. Horn, S. Eliasson, L. J. Jensen, DistiLD database: diseases and traits in linkage disequilibrium blocks, Nucleic Acids Research 40 (D1) (2011) D1036–D1040. doi:10.1093/nar/gkr899.

[22] J. J. Yang, D. Grissa, C. G. Lambert, C. G. Bologa, S. L. Mathias, A. Waller, D. J. Wild, L. J. Jensen, T. I. Oprea, TIGA: Target illumination GWAS analytics, Bioinformatics 37 (21) (2021) 3865–3873. doi:10.1093/bioinformatics/btab427.

[23] N. Rappaport, M. Twik, I. Plaschkes, R. Nudel, T. Iny Stein, J. Levitt, M. Gershoni, C. P. Morrey, M. Safran, D. Lancet, MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search, Nucleic Acids Research 45 (D1) (2016) D877–D887. doi:10.1093/nar/gkw1012.

[24] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, L. I. Furlong, The DisGeNET knowledge platform for disease genomics: 2019 update, Nucleic Acids Research 48 (D1) (2019) D845–D855. doi:10.1093/nar/gkz1021.

[25] T. K. Sheils, S. L. Mathias, K. J. Kelleher, V. B. Siramshetty, et al., TCRD and Pharos 2021: mining the human proteome for disease biology, Nucleic Acids Research 49 (D1) (2020) D1334–D1346. doi:10.1093/nar/gkaa993.

[26] D. Ochoa, A. Hercules, M. Carmona, et al., Open Targets Platform: supporting systematic drug-target identification and prioritisation, Nucleic acids research 49 (D1) (2021) D1302–D1310. doi:10.1093/nar/gkaa1027.

[27] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J.

Jensen, C. v. Mering, STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, Nucleic Acids Research 47 (D1) (2018) D607–D613. doi:10.1093/nar/gky1131.

[28] L. M. Schriml, E. Mitraka, J. Munro, B. Tauber, M. Schor, L. Nickle, et al., Human Disease Ontology 2018 update: classification, content and workflow expansion, Nucleic Acids Research 47 (D1) (2018) D955–D962. doi:10.1093/nar/gky1032.

[29] K. C. Nastou, G. I. Nasi, P. L. Tsiolaki, Z. I. Litou, V. A. Iconomidou, AmyCo: the amyloidoses collection, Amyloid 26 (3) (2019) 112–117. doi:10.1080/13506129.2019.1603143.

[30] B. I. Hutchins, X. Yuan, J. M. Anderson, G. M. Santangelo, Relative Citation Ratio (RCR): A new metric that uses citation rates to measure influence at the article level, PLOS Biology 14 (9) (2016) 1–25. doi:10.1371/journal.pbio.1002541.

[31] R. I. Doğan, W. J. Wilbur, D. C. Comeau, BioC and simplified use of the PMC open access dataset for biomedical text mining, in: Proceedings of the 4th Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, 2014.

[32] D. S. Chawla, A single 'paper mill' appears to have churned out 400 papers, sleuths find, Sciencedoi:10.1126/science.abb4930.

[33] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification (2016).

[34] P. Flicek, I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, al., Ensembl 2013, Nucleic Acids Research 41 (D1) (2012) D48–D55. doi:10.1093/nar/gks1236.

[35] K. A. Gray, L. C. Daugherty, S. M. Gordon, R. L. Seal, M. W. Wright, E. A. Bruford, Genenames.org: the HGNC resources in 2013, Nucleic Acids Research 41 (D1) (2012) D545–D552. doi:10.1093/nar/gks1066.

[36] E. Pafilis, S. Pletscher-Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, L. J. Jensen, The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text., PLOS ONE 8 (6). doi:10.1371/journal.pone.0065390.

[37] H. Else, R. Van Noorden, The fight against fake-paper factories that churn out sham science, Nature 591 (7851) (2021) 516–519. doi:10.1038/d41586-021-00733-5.

[38] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, et al., The GeneCards suite: From gene data mining to disease genome sequence analyses, Current Protocols in Bioinformatics 54 (1) (2016) 1.30.1–1.30.33. doi:https://doi.org/10.1002/cpbi.5.

[39] D. C. Cannon, J. J. Yang, S. L. Mathias, et al., TIN-X: target importance and novelty explorer, Bioinformatics (Oxford, England) 33 (16) (2017) 2601–2603. doi:10.1093/bioinformatics/btx200.

[40] A. Lachmann, B. M. Schilder, M. L. Wojciechowicz, D. Torre, M. V. Kuleshov, A. B. Keenan, A. Ma'ayan, Geneshot: search engine for ranking genes from arbitrary text queries, Nucleic acids research 47 (W1) (2019) W571–W577. doi:10.1093/nar/gkz393.

[41] A. D. Rouillard, G. W. Gundersen, N. F. Fernandez, Z. Wang, C. D. Monteiro, M. G. McDermott, A. Ma'ayan, The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins, Database : the journal of biological databases and curation 2016. doi:10.1093/database/baw100.

[42] N. T. Doncheva, J. H. Morris, J. Gorodkin, L. J. Jensen, Cytoscape StringApp: Network analysis and visualization of proteomics data, Journal of Proteome Research 18 (2) (2019) 623–632. doi:10.1021/acs.jproteome.8b00702.

Response to reviewers (Discussion)

The comments given by the reviewers of our manuscript are referred to in the following manner: R1-C1 indicates the first comment of Reviewer1.

Reviewer 1:

We sincerely thank you for your positive feedback on our paper and for your careful review.

Reviewer 2:

We really appreciate your careful reading and we thank you for your thoughtful and constructive comments, which helped improve the manuscript.. Below, we have addressed all of them.

R2-C1: *Besides for the [tagger](https://github.com/larsjuhljensen/tagger), the code to generate DISEASES does not appear to be available. This makes it hard to get a detailed look at the methods beyond what is described in the manuscript. It also prevents the community from contributing and creating alternative datasets that reuse parts of the DISEASES infrastructure. It's not clear what the reason is for keeping the code private.*

Response:

As mentioned by the reviewer, the tagger is indeed available. The same is true for the entire Python codebase that runs the website (https://github.com/larsjuhljensen/mamba). However, some other parts are not for a variety of reasons. Firstly, the other code bases are designed specifically to populate the backend database within our local server setup. It would not be easy to set up a clone of it elsewhere, and we would not have the resources to support anyone attempting to do so. The code would thus be provided with limited documentation, as-is, and with no support at all. We do not believe that would be of much use to others, and it would more likely than not lead to angry people, when emails went unanswered.

There is also a second issue: the code used to make parts of the DISEASES database is a shared codebase for several other resources, including COMPARTMENTS, TISSUES and STRING. Parts of the code in the repository were thus developed as part of other projects by people who are not authors on this paper (or on the earlier DISEASES paper), and parts of the code is not even relevant to DISEASES, but is purely used by these other resources. Making it available under an open source license would thus not be possible for us, as we are not the owners of the code. It would thus require tracking down who made every single part of the code in the repository and obtaining their permission first.

If people have detailed questions for how something was done, we are happy to provide them a copy of the relevant script. But making the entire codebase publicly available is not feasible.

R2-C2: *If the resource is regenerated weekly using continuous integration, it would also be valuable to give the community read access to the logs. This would occur automatically if the resource is built using a service like GitHub Actions. It's not clear from the manuscript how or where the weekly builds occur.*

Response:

The weekly builds occur on an in-house server, which runs update shell scripts as cron jobs. These update scripts in turn run things such as the tagger, which is coded in C++. This plus the sheer amount of data being processed in the text mining would make it very difficult to migrate to e.g. GitHub actions.

R2-C3: Versioned permalinks would be extremely helpful, so users can pin their analyses to old versions of the data without having to archive it themselves.

Response:

Making versioned permalinks would be a very inefficient use of resources. The reason is that we update the entire database every weekend and that the text-mining scores for almost all associations will change a bit due to the addition of new abstracts to PubMed. We would thus have to archive a full version of the entire database every week, knowing that the vast majority of these archived versions would not be the basis of any analysis. To not unnecessarily waste resources, we thus believe that it is better that users archive the specific versions that they actually used rather than us archiving every version ever made.

R2-C4: *On the [downloads page] (https://web.archive.org/web/20220110224835/https://diseases.jensenlab.org/Downloads), the Figshare [archive link](https://figshare.com/authors/Lars_Juhl_Jensen/96428) is to a user profile, and not an archive of DISEASES v2. On the user page, there are multiple DISEASES v1 records, but the sigle [DISEASES v2] (https://figshare.com/articles/dataset/DISEASES_v2_gold_standard_/17075708) record contains only the gold standard data.*
Ideally, there would be a Figshare record for DISEASES 2.0 (or multiple records for different components, e.g. tagging dictionaries might be their own record.) Rather than create a new record for each DISEASES v2 release, it would be best to create multiple versions within the same record.

Response: This is a good point of using a Figshare record that includes all the components related to DISEASES 2.0.

We have added a Figshare record for DISEASES 2.0 that includes the following components: text mining, knowledge, experiment, integrated score and dictionaries

R2-C5: The downloads page links to, but not to `human_dictionary.tar.gz` with the gene dictionary. Perhaps it'd be less confusing to just link to , which already contains links to all the predefined dictionaries?

Response: The intent was that the diseases_dictionary.tar.gz file should be self-contained and have both disease and human gene names in it. However, it only contained disease names. We have now fixed this.

R2-C6: *I am left wondering what the difference between `diseases_tagger.tar.gz` is compared to the software from the GitHub repository.*

Response: The DISEASES tagger is a specific version of the tagger software allowing users to easily run named-entity recognition on a directory of text files. It is based on the full-fledged tagger available from GitHub, which is intended to be run on much bigger corpora and have more functionality, but also requires more work to prepare the corpus in the specific format required to process it efficiently (i.e. each document is not a separate file).

We have edited the DISEASES website to also link to the full version of tagger.

R2-C7: *The association score TSV downloads don't contain column names as the first row, nor are the columns documented on the downloads page. It's best to keep column names with the data.*

Response: We agree that it would have been a good idea to have column names in the files from the start. However, at this stage it is too late to add them, since many people have workflows that we risk breaking by changing the file format now. We have instead added the format to the downloads page and as a documentation file on Figshare.

R2-C8: *`excluded_documents.txt` provides a list of 826 publications from paper mills that should be excluded by T&DM approaches. This is a useful resource beyond this project. Is the source of this file tracked with collaborative version control? It could be beneficial to have this file in a GitHub repo where the community could suggest additional exclusions in the future.*

Response: We fully agree that this file would be useful for text-mining projects in general, which is precisely why we make it available. To the best of our knowledge, there is not a version-controlled community resource that gathers all the paper mill publications that have been discovered to date. We very much wish that there would be one, since it would make our work much easier — the list of 826 publications was made by combining lists from several different spreadsheets shared by different people in the community, some of which had PubMed IDs and others DOIs that we had to map. However, since we are not involved in discovering the paper mills, it would really not be appropriate for us to be the ones managing such a resource and thereby be the gatekeepers. However, we do have our own file in version control, so if there were ever the need to go back or see when a certain publication was added to the list, we can do that.

***R2-C9:** Many users might be interested in an integrated score that combines the knowledge, experiments, and text mining channels. We've discussed methods for this [in the past] (https://think-lab.github.io/d/91/#7). DISEASES v1 included an integrated score ([archived here] (https://github.com/dhimmel/diseases/blob/60f955165e297caaba86d2f3d370fba71d3b26bc/download/human_disease_integrated_full.tsv.gz)), but it looks like DISEASES v2 removes this score. What is the reason for the removal of the integrated score?*

Response: We agree that having an integrated score that combines all the different types of evidence can be useful. Like DISEASES v1, DISEASES v2 is not showing or using such a score anywhere in the web resource; however, it does exist in the backend database and we do have a download file with it. The reason why we do not show it, use it, or mention it in the manuscript is that the way scores are combined is very ad hoc; while the scores may be useful, we have no way to scientifically justify that the way we combine confidence scores is the right way to do it. It thus should not be thought of as a feature of the DISEASES database or something that is part of a scientific publication. That being said, we are happy to share the scores if people nonetheless find them useful.

As a compromise, we have thus added a link on the Downloads to the already existing file with the integrated scores, but have marked it as an “experimental” feature to warn users that it is not something they should rely on.

***R2-C10:** The methods for producing the filtered datasets that "contain only the non-redundant associations" are not clear to me. Is this related to backtracking of evidence to parent diseases?*

Response:

This is difficult to explain briefly beyond what was already stated on the download page. It relates to backtracking, but it is not generally as simple as not including parent terms.

The filtered files are available for download to allow other resources, notably GeneCards, to get a non-redundant list of diseases for a gene of interest, which is consistent with what is shown in the DISEASES web interface.

For the knowledge channel, we have the explicitly annotated terms and infer parent terms from those. In this case, the filtered set is indeed simply the explicitly annotated terms, thus excluding the inferred parents.

However, especially for text mining, it is a lot more complex. If you look at a gene like LRRK2, a famous gene in Parkinson’s disease, it will of course be mentioned a lot in the literature with this disease. And these co-occurrences will be backtracked to parent terms like neurodegenerative disease. However, LRRK2 will also explicitly co-occur a lot in the literature with the term neurodegenerative disease, since many papers about Parkinson’s disease will mention that it is a neurodegenerative disease. One can thus not simply filter out inferred parent terms to fix this problem. We thus need to make a clever decision about which term to show: Parkinson’s disease or neurodegenerative disease. Our solution is to show the highest scoring term, exclude all its parents and children, show the second-highest scoring (not excluded) term, exclude all its parents and children, etc.

Again, the filtered purely exist to be able to provide a non-redundant summary of the diseases for a gene of interest. Anyone who wants to do systematic analysis of the gene–disease network or wants to get the list of all genes relevant to a disease should use the full files. Given how niche the use case for the filtered files is and how much text it takes to explain how precisely they are made and why, we do not believe it makes sense to go into this in the manuscript.

We are considering removing the links to these files from the download page, because they cause confusion and 99% users should not use them. Instead we would purely provide these files to the small number of users (e.g. GeneCards) for whom they are intended.

***R2-C11:** There's a missing word in the following sentence: "The Experiments channel has **CHANGED** in many ways between the two versions".*

Response: Thanks for spotting this — we have added the missing word.

Reviewer 3:

Thank you for going through our work with great interest. We really appreciate your feedback on our paper. Below, we have addressed your comments.

***R3-C1:** Indeed, a brief perusal of the resource for “Alzheimer's disease [DOID:10652]”, a well-studied disease, showed that there was a good agreement between the “Evidence” and the “Text Mining” tracks. However, the top genes in the “Experimental track” did not include any of the well-known genes associated with the disease (e.g. APOE - https://www.nature.com/articles/mp201152, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6463297/, the stronger genetic risk for AD). The reason could be that APOE’s signal is so strong that it is often removed from GWAS analyses. This leads me to wonder about the quality of the experimental channel in this case and similar, or the validity of GWAS for generating gene based annotation. I think a ROC curve for the experimental channel, as well as an overlap diagram, looking at the comparison between the experimental channel and the evidence channel would be an important addition, with the associated caveats.*

Response: Thank you for looking into this specific disease and noticing this problem. As GWAS-based associations are imported into DISEASES from the TIGA database, it is a problem that would have to be solved in that resource. The overall quality of TIGA gene—disease associations was already evaluated and published recently in the TIGA paper https://academic.oup.com/bioinformatics/article-abstract/37/21/3865/6292081, which includes the ROC curve suggested by the reviewer and yielded an AUC of 0.73. Adding an overlap diagram between the experimental channel and the knowledge channel would require choosing an arbitrary confidence cutoff and would be redundant with the ROC curve, as the gold standard is based on the knowledge channel.

We have added a short section mentioning this to the end of Materials and Methods.

R3-C2: *I could not find an explanation for the calculation of the confidence score, in the paper or the resource itself. For example – what is the strategy when there's conflicting evidence from curated data vs Text-mined data (meaning the disease gene pair exists in both, but not the association). Since this is an update, it doesn't have to be in the paper itself, but I'd expect for it to be mentioned on the website.*

Response: Given a disease--gene pair, the confidence score is computed for every channel independently, and is comparable across evidence types. A detailed description of the confidence scores for text mining is given in the original publication of DISEASES.

We have added a section to the end of Materials and Methods, which briefly summarizes how confidence scores for text mining are calculated and, more importantly, describe how comparable confidence scores for TIGA were derived (which was not described in the earlier papers.

R3-C3:*The paper does not discuss the meaning of "associated". Are they trying to aggregate causal genes ? Genes that change in expression? Genes that are drug targets ? Any gene ? This needs to at least be discussed, even if the goal is to cast a wide net to serve exploratory analyses.*

Response: While it is always dangerous to claim causality, the intended scope of DISEASES is causal associations between genes and diseases, without limiting it to drug targets. The gold standard that we specifically consists of manually curated genetic associations that are believed to be causal, and as such our confidence scores reflect that. This is also the reason why we do not include expression data as evidence in the experiments channel. However, since we have confidence scores, we obviously do not expect all associations in DISEASES to be true causal associations.

We have added a couple of sentences on this to the first section of Results & Discussion.

R3-C4: *The focus on AmyCo is unclear. It seems that resources could be put to a better use by including broader databases rather than this specific class of conditions. I wonder what led the authors to choose this specific one, over, for example Orphanet, which contains a large set of gene disease associations, and is also freely available.*

Response: We fully agree that AmyCo is a fairly narrow database, covering only amyloidoses. It was added to DISEASES as part of a specific research project focused on these diseases, and as it contains high quality manually curated associations, we see no reason to remove it. Regarding Orphanet, it is indeed a resource that we are considering to add to DISEASES. However, it would take a significant amount of work to combine Disease Ontology and Orphanet, since some diseases in Orphanet have Disease Ontology terms whereas most do not. One would thus have to assign all Orphanet diseases that are not already in Disease Ontology as child terms of the appropriate Disease Ontology term. That work is simply beyond the scope of this release of the database and would require dedicated funding for the project given the very large number of diseases in Orphanet.

We now mention in the text that AmyCo database is a manually curated database.