HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

DATABASE

GAAD: A Gene and Autoimmiune Disease Association Database

Guanting Lu [1,#,a], Xiaowen Hao [2,#,b], Wei-Hua Chen [2,*,c], Shijie Mu [1,*,d]

[1] Department of Blood Transfusion, Tangdu Hospital, Fourth Military Medical University, Xi'an 710032, China
[2] MOE Key Laboratory of Molecular Biophysics, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

Received 3 December 2017; revised 4 April 2018; accepted 22 May 2018
Available online 27 September 2018

Handled by Quan-Zhen Li

KEYWORDS

Autoimmune diseases;
Disease–gene association;
Database;
Text mining

Abstract **Autoimmune diseases** (ADs) arise from an abnormal immune response of the body against substances and tissues normally present in the body. More than a hundred of ADs have been described in the literature so far. Although their etiology remains largely unclear, various types of ADs tend to share more associated genes with other types of ADs than with non-AD types. Here we present GAAD, a gene and AD association **database**. In GAAD, we collected 44,762 associations between 49 ADs and 4249 genes from public databases and MEDLINE documents. We manually verified the associations to ensure the quality and credibility. We reconstructed and recapitulated the relationships among ADs using their shared genes, which further validated the quality of our data. We also provided a list of significantly co-occurring gene pairs among ADs; with embedded tools, users can query gene co-occurrences and construct customized co-occurrence network with genes of interest. To make GAAD more straightforward to experimental biologists and medical scientists, we extracted additional information describing the associations through **text mining**, including the putative diagnostic value of the associations, type and position of gene polymorphisms, expression changes of implicated genes, as well as the phenotypical consequences, and grouped the associations accordingly. GAAD is freely available at http://gaad.medgenius.info.

[*] Corresponding authors.
E-mail: weihuachen@hust.edu.cn (Chen WH), musj1963@fmmu.edu.cn (Mu S).
[#] Equal contribution.
[a] ORCID: 0000-0002-8556-8687.
[b] ORCID: 0000-0002-0175-1186.
[c] ORCID: 0000-0001-5160-4398.
[d] ORCID: 0000-0003-1267-6000.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2018.05.001

Introduction

An autoimmune disease (AD) is a condition that arises from an abnormal immune response to a normal body part [1]. ADs may be restricted to certain organs or involve a particular tissue in different places. For example, autoimmune thyroiditis only affects the thyroid glands [2], while Goodpasture's disease can affect the basement membrane in both the lung and the kidney [3]. Many ADs, with the exception of rheumatoid arthritis (RA) and autoimmune thyroiditis, are individually rare; but together they affect more than 10% of the population worldwide, with annual net increase of ∼19% in incidents [4]. It has been estimated in the United States alone that up to 23.5 million Americans suffer from in total 24 ADs, accounting for ∼7.5% of the whole population (see also https://www.aarda.org, the American Autoimmune Related Diseases Association for the latest statistics), and the annual health care costs for ADs are in the range of $100 billion (data from https://www.womenshealth.gov/a-z-topics/autoimmune-diseases; accessed on July 19, 2017). In comparison, cancer affects up to 9 million Americans and costs ∼$57 billion (data source: American Cancer Association; accessed on July 19, 2017).

Similar to other complex diseases such as cancer, many genes have been reported to be associated with ADs, but together they only account for a minor proportion of the disease risks. For example, a recent fine-mapping study using 33,595 patients with inflammatory bowel disease (IBD) estimates that ∼22%–28% of the risks can be explained by in total 94 genomic regions [5]. Previous observations in monogenetic twins indicated that genetic influences only account for ∼25%–40% of the disease risks [6]. These results indicate that either there are hidden factors such as epigenetics and/or protein post-translational modifications (PTMs) that have not been included in their experimental design, or environmental influences and/or gene–environment interactions constitute the predominant factor for disease risks. For example, it has been shown that the gut microbiota–host interactions have been reported to significantly shape the genetic architecture of some ADs, such as IBD [7], lupus nephritis [8], RA [9,10], multiple sclerosis [11], and Graves' disease [12]. Nonetheless, due to the diverse and difficult-to-quantify nature of environmental factors, genetic background and DNA mutations are still the major measurable disease risks for ADs so far.

Although the triggering factors for many ADs remain unknown [6], genome-wide association studies (GWAS) have identified many nucleotide polymorphisms, with some of them likely to be causal, and associated risk loci [5,7]. Not surprisingly, many ADs share associated risk loci with other types of ADs significantly more than with other disease types [13], due to their shared triggering factors, and many ADs could co-occur in the same patients [14,15]. In addition, risk loci in ADs often act independently, with an additive effect on the disease outcomes [7].

Disease-specific databases are useful for researchers to find and retrieve the appropriate information. In recent years, the number of biomedical databases has increased dramatically. However, a gene–disease association database dedicated to ADs is still unavailable. In this study, we presented GAAD, a gene–disease association database for ADs, aimed to enhance our understanding of ADs. The following information is collected into a central database: (i) a comprehensive gene catalog associated with ADs, facilitating genome-wide analyses as well as gene-centered biological studies; (ii) relevant features describing the characteristics of such associations, including the putative diagnostic values of the association, time of disease onset due to dysfunction of an associated gene and expression changes of causal and/or implicated genes during disease progression, allowing quick selection of suitable biomarkers by experimental and/or clinical researchers. In addition, we also obtain a disease network using co-occurring gene pairs commutated from our collected data, allowing researchers to compare disease similarity and get hints on novel causal genes of ADs. For example, if most of the known causal genes of an AD are associated with another AD, then other genes associated with the latter disease could also be associated with the former disease.

In GAAD, we collected in total 44,763 associations between 4249 genes and 49 Ads, each association as an independent piece of evidence supporting the link between a gene and a disease. We obtained the association from public databases including the NCBI Gene database and GeneCards [16], literatures describing large-scale GWAS analyses [17], and text mining results from MEDLINE documents. We manually verified the associations to ensure the data quality. Additionally, we extracted available information describing the associations through text mining, including the putative diagnostic value of the associations, type and position of gene polymorphisms, expression changes of implicated genes, and the phenotypical consequences, and grouped the associations accordingly, allowing experimental biologists and medical scientists to quickly and easily find relevant information to genes of interest.

Implementation

GAAD is implemented with Javascript, HTML, PHP, and AngularJS (a model-view-controller frame work for creating Javascript web applications; https://angularjs.org). GAAD uses Bootstrap (http://getbootstrap.com) for most of its front-end (*i.e.*, the user interface) components and adopts MySQL (https://www.mysql.com) as relational database management system to store all the data. All codes are developed using PhpStorm (https://www.jetbrains.com/phpstorm/) Educational version. The whole website is hosted on a CentOS (one of the most popular Linux distributions; https://www.centos.org) operating system.

Database content and usage

Disease selection

We took the disease list from the American Autoimmune Related Diseases Association (www.aarda.org), cross-referenced with the list of diseases included in NCBI. Consequently, 49 ADs, for which gene associations can be found by our data collection process (see below), were retained.

Data acquisition

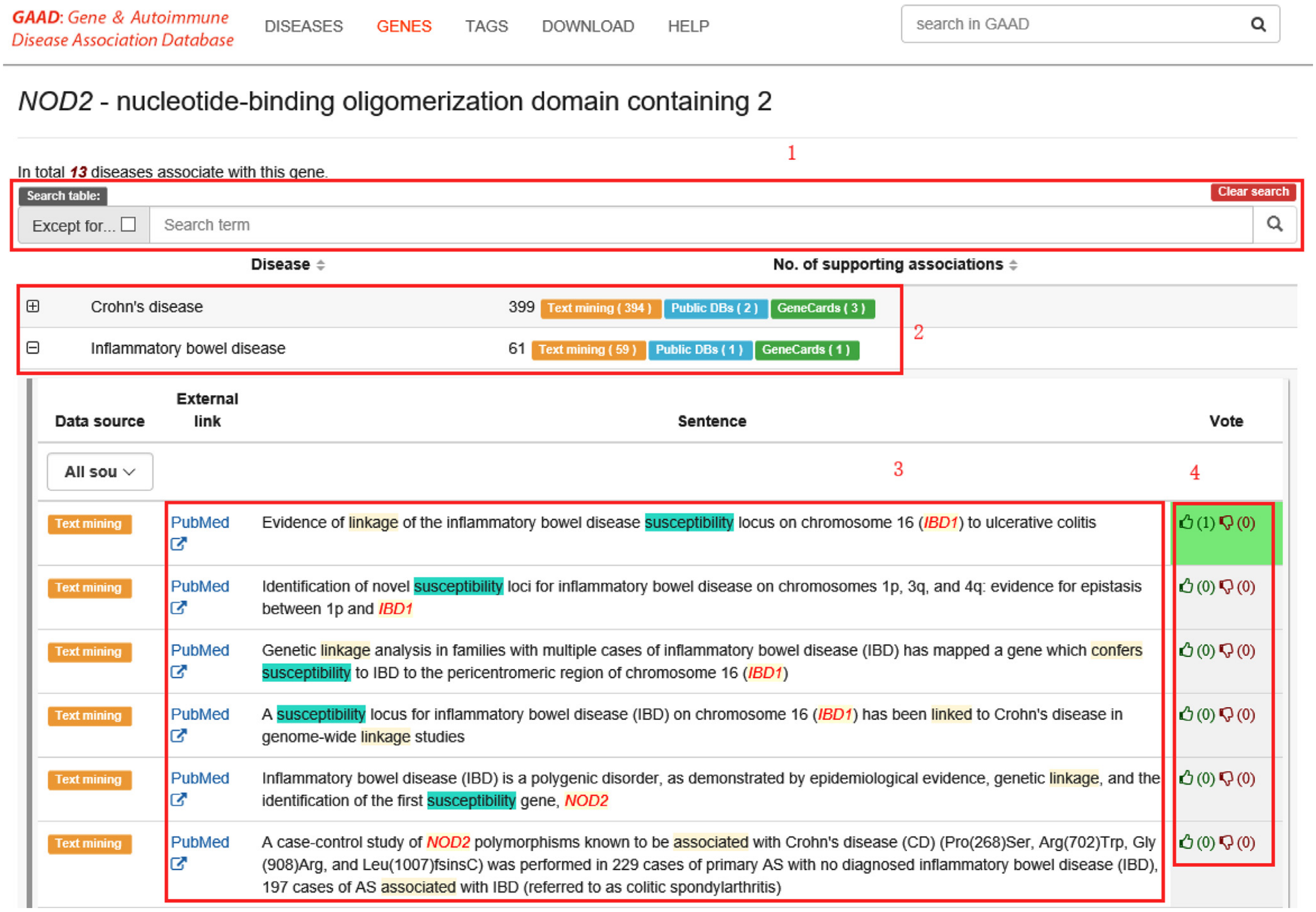
We performed text mining analysis using a customized Python pipeline by searching for co-occurrences of genes (*e.g.*, gene symbol, locus ID, common name, alias, and variant) and diseases of interest in the titles or abstracts of MEDLINE documents available from the NCBI PubMed database (https://www.ncbi.nlm.nih.gov/pubmed). The criteria include: (1) the gene name and the disease name should be in the same sentence; (2) this sentence should also contain one of the following key words/phrases or their aliases: *aberrant*, *account for*, *altered*, *associated*, *caused*, *confer*, *contribute*, *curb*, *downregulate*, *dysregulated*, *elevate*, *evoked*, *higher*, *implicated*, *increase*, *induce*, *influence*, *interact*, *involved*, *lead to*, *link*, *mediate*, *modulate*, *overexpressed*, *prevent*, *protective*, *reduce*, *regulate*, *related*, *relationship*, *treat*, *biomarker*, *clinic trial*, *therapy*, *diagnose*, *risk factor*, *target*, *prognostic*, *pathogenic*, *pathogenesis*, *progression*, and *predisposing factor*. In addition, we also extracted features describing the relationships between genes and ADs (see above) from the identified sequences. We obtained 135,673 associations from 90,918 MEDLINE documents through initial text mining. The retrieved information was further verified by three rounds of manual inspection.

We also searched for gene–disease associations in the NCBI Gene database using disease names (and aliases) as input. The retrieved entries were also manually curated and incorporated into GAAD.

In addition, we used an in-house Perl script to search and retrieve gene–disease associations from the GeneCards database (http://www.genecards.org), which uses a disease relevance score to describe the credit of each association. The relevance score of an association was calculated by comparing "the observed number of MEDLINE documents where both elements appear together and the number of MEDLINE documents where both appear independently" with an expected value based on a hypergeometric distribution, which reflects an order of importance, with higher disease relevance score indicating more co-occurrences.

Data analyses

We calculated gene co-occurrences for all possible pairs using Fisher's exact test to count data (the fisher.test() function)



NOD2 - nucleotide-binding oligomerization domain containing 2

In total ***13*** diseases associate with this gene.

	Disease	No. of supporting associations
⊞	Crohn's disease	399 Text mining (394) Public DBs (2) GeneCards (3)
⊟	Inflammatory bowel disease	61 Text mining (59) Public DBs (1) GeneCards (1)

Data source	External link	Sentence	Vote
Text mining	PubMed ↗	Evidence of linkage of the inflammatory bowel disease susceptibility locus on chromosome 16 (*IBD1*) to ulcerative colitis	👍 (1) 👎 (0)
Text mining	PubMed ↗	Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for epistasis between 1p and *IBD1*	👍 (0) 👎 (0)
Text mining	PubMed ↗	Genetic linkage analysis in families with multiple cases of inflammatory bowel disease (IBD) has mapped a gene which confers susceptibility to IBD to the pericentromeric region of chromosome 16 (*IBD1*)	👍 (0) 👎 (0)
Text mining	PubMed ↗	A susceptibility locus for inflammatory bowel disease (IBD) on chromosome 16 (*IBD1*) has been linked to Crohn's disease in genome-wide linkage studies	👍 (0) 👎 (0)
Text mining	PubMed ↗	Inflammatory bowel disease (IBD) is a polygenic disorder, as demonstrated by epidemiological evidence, genetic linkage, and the identification of the first susceptibility gene, *NOD2*	👍 (0) 👎 (0)
Text mining	PubMed ↗	A case-control study of *NOD2* polymorphisms known to be associated with Crohn's disease (CD) (Pro(268)Ser, Arg(702)Trp, Gly (908)Arg, and Leu(1007)fsinsC) was performed in 229 cases of primary AS with no diagnosed inflammatory bowel disease (IBD), 197 cases of AS associated with IBD (referred to as colitic spondylarthritis)	👍 (0) 👎 (0)

Figure 1 The "GENES" page for gene *NOD2* and its associated diseases

Essential elements of this page include: (1) widgets allowing users to search for or filter out diseases by user-supplied search terms; (2) a list of associated diseases and supporting evidence. The supporting evidence is hidden by default but can be displayed by clicking the "+" sign on the left; (3) text mining results with different types of keywords highlighted in different font colors and background colors; and (4) widgets allowing users to vote on the quality of the corresponding text mining entry. The color of the background cell can be automatically changed according to the voting (*i.e.*, gray if there are equal numbers of "up" and "down" votes; green if there are more "up" votes; and red if there are more "down" votes). See http://gaad.medgenius.info/genes/NOD2 for details.

Table 1 Top 10 autoimmune diseases ranked according to the total number of associated genes

Autoimmune disease	No. of associated genes
Rheumatoid arthritis	1324
Multiple sclerosis	977
Systemic lupus erythematosus	876
Inflammatory bowel disease	843
Crohn's disease	788
Psoriasis	709
Ulcerative colitis	680
Parkinson's disease	475
Systemic sclerosis	469
Atopic dermatitis	394

Note: A complete and up-to-date list of autoimmune diseases and the number of associated genes is available at http://gaad.medgenius.info/diseases/.

implemented in R (https://cran.r-project.org). This R function takes a 2 × 2 contingency table as input; the data include four values which represent the numbers of ADs associated with (1) both genes in the pair, (2) only one gene, (3) only the other gene and (4) none of them. We reported significantly co-occurring gene pairs in GAAD.

For the two genes in a co-occurring pair, we also tested whether they are in the same KEGG pathways using R modules including pathview [18], KEGGREST (an R package utilizing KEGG Representational State Transfer (REST) programmable interface to access KEGG data; available at https://bioconductor.org/packages/release/bioc/html/KEGGREST.html), and gage (https://bioconductor.org/packages/release/bioc/html/gage.html). Genes in the same pathways are likely to have similar and/or relevant functions. However, it does not mean that two genes are not functionally similar if they are not in the same pathway, since only small fractions of human genes (6356 in this case) are involved in various pathways.

To test whether co-occurring genes are more likely to share similar functions, we also performed an additional analysis using Gene Ontology (GO), since GO covers more human proteins/genes than the KEGG database. We used a R package GOSemSim [19] to perform such analysis. GOSemSim computes semantic similarity among GO terms and reports a similarity score for each input gene pair ranging from 0 to 1, where 0 indicates no similarity and 1 indicates at least one GO term is shared by two genes (although it does not mean their functions are identical).

Database contents and organization

The main contents of our database are organized into two pages, namely the "DISEASES" page and the "GENES" page.

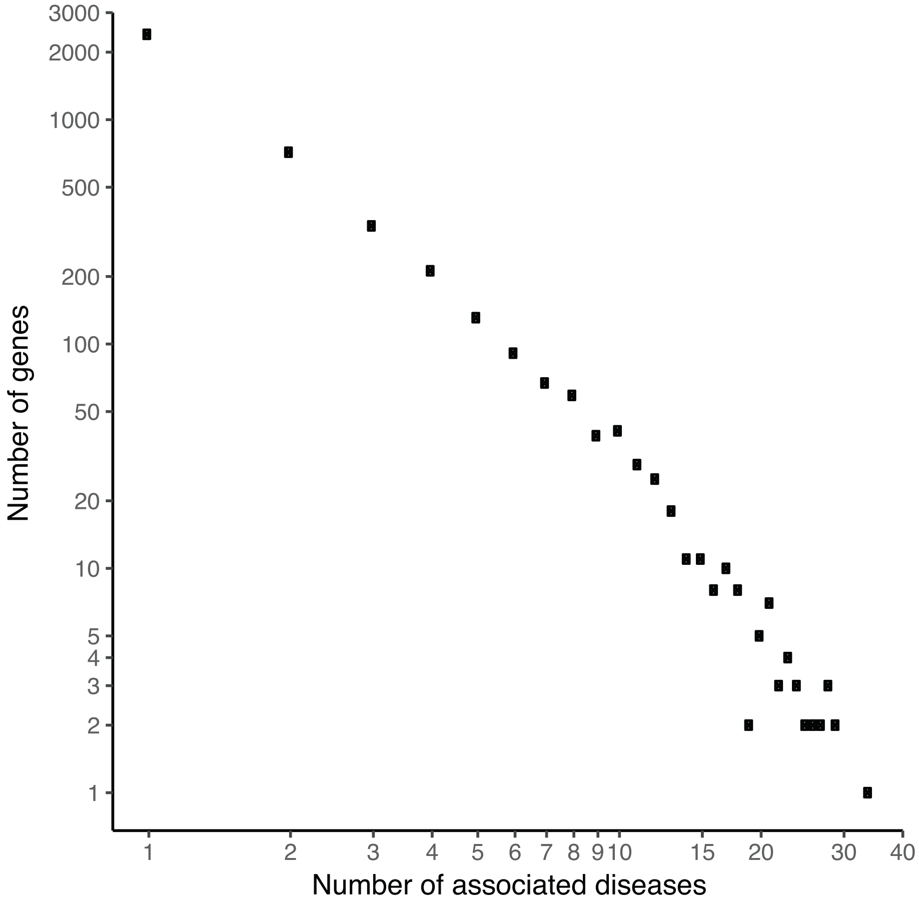


Figure 2 The number of associated diseases for the 4249 genes follows a power-law distribution
Overall the distribution of the number of diseases that a gene is associated with follows the power-law distribution. A large number of genes are associated with a few diseases, while a few genes are associated with a large number of diseases. Values on both axes are log-transformed.

At the beginning of "DISEASES" ("GENES") page, a summary table lists available diseases (genes) and the number of associated genes (diseases). We also provide widgets that allow users to search for or filter out diseases (genes) by user-supplied keywords. After clicking on a disease (gene) name, users will be redirected to a new page containing detailed information for the disease (gene), including a list of associated genes (diseases) and detailed evidence supporting the associations.

Most of our supporting evidence was obtained through text mining, we thus include the sentences in which the disease–gene association was described in the evidence view. We also highlight different types of keywords with different font and background colors (**Figure 1**). Text mining results are often error-prone; we thus carried out multi rounds of aggressive manual curation to remove any possible false results. However, despite these efforts, there might still be small fraction of false positive results. We thus include a user-feedback mechanism, allowing users to vote on each text mining entry. As shown in Figure 1, users can use the thumb-up and thumb-down buttons to vote on the creditability of the entry; the background color of the table cell will be changed automatically according to the voting results.

We collected in total 41,235 associations from 19,299 MEDLINE documents through text mining followed by multi rounds of manual curation (see Data acquisition). We obtained additional 3489 associations from public databases including NCBI Gene and GeneCards. Together, we collected in total 44,762 associations between the 49 ADs and 4249 genes in our database.

The number of genes associated with an AD ranges from 1324 (RA) to 2 (vitiligo-associated multiple autoimmune disease), as shown in **Table 1**. Out of the associated genes in GAAD, more than half of the genes are associated with only one disease. Overall, the distribution of the number of diseases which a gene is associated with follows the power-law distribution (data from https://en.wikipedia.org/wiki/Power_law; accessed on July 20, 2017) (**Figure 2**). In total 33 genes are associated with more than 20 ADs (**Table 2**); not surprisingly, most genes are involved in immune response.

It should be pointed out that in this study we adopted a broad definition of "gene", which refers to not only any discrete locus of heritable, genomic sequence that affect an organism's traits, but also its products including RNAs and proteins.

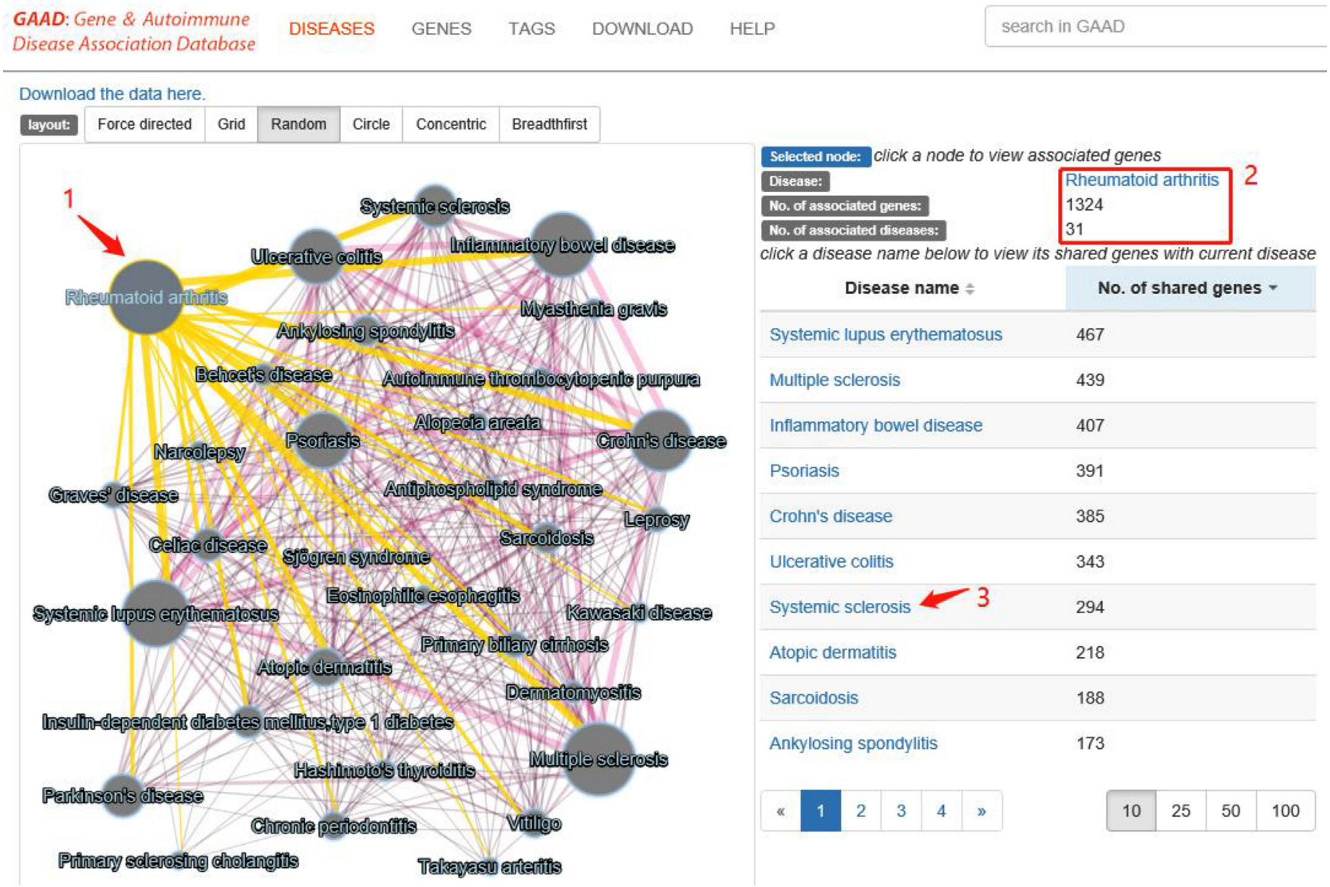
Genes shared by ADs and significantly co-occurring gene pairs

ADs collected in our database share significant proportion of their associated genes with other types of ADs. For example, for any given AD, it shares at least one associated gene with another AD. Therefore a network based on shared genes

Table 2 Genes associated with > 20 autoimmune diseases in GAAD

Gene ID in GAAD	No. of associated diseases	Gene symbol	Protein encoded
7124	34	*TNF*	Tumor necrosis factor
3569	29	*IL6*	Interleukin 6
920	29	*CD4*	CD4 molecule
3586	28	*IL10*	Interleukin 10
3558	28	*IL2*	Interleukin 2
3123	28	*HLA-DRB1*	Major histocompatibility complex, class II, DR beta 1
3605	27	*IL17A*	Interleukin 17A
3458	27	*IFNG*	Interferon, gamma
50943	26	*FOXP3*	Forkhead box P3
3565	26	*IL4*	Interleukin 4
3627	25	*CXCL10*	Chemokine (C-X-C motif) ligand 10
3553	25	*IL1B*	Interleukin 1, beta
51497	24	*NELFCD*	Negative elongation factor complex member C/D
1493	24	*CTLA4*	Cytotoxic T-lymphocyte-associated protein 4
925	24	*CD8A*	CD8a molecule
7422	23	*VEGFA*	Vascular endothelial growth factor A
7099	23	*TLR4*	Toll-like receptor 4
7040	23	*TGFB1*	Transforming growth factor, beta 1
3606	23	*IL18*	Interleukin 18
51561	22	*IL23A*	Interleukin 23, alpha subunit p19
3576	22	*CXCL8*	Chemokine (C-X-C motif) ligand 8
1401	22	*CRP*	C-reactive protein, pentraxin-related
6347	21	*CCL2*	Chemokine (C-C motif) ligand 2
4524	21	*MTHFR*	Methylenetetrahydrofolate reductase (NAD(P)H)
4282	21	*MIF*	Macrophage migration inhibitory factor (glycosylation-inhibiting factor)
3630	21	*INS*	Insulin
3119	21	*HLA-DQB1*	Major histocompatibility complex, class II, DQ beta 1
3106	21	*HLA-B*	Major histocompatibility complex, class I, B
720	21	*C4A*	Complement component 4A (Rodgers blood group)
90865	20	*IL33*	Interleukin 33
50616	20	*IL22*	Interleukin 22
3559	20	*IL2RA*	Interleukin 2 receptor, alpha
3383	20	*ICAM1*	Intercellular adhesion molecule 1
959	20	*CD40LG*	CD40 ligand

Note: A complete and up-to-date list of genes associated with > 20 autoimmune diseases is available at http://gaad.medgenius.info/genes/.



Selected node: click a node to view associated genes

Disease: Rheumatoid arthritis

No. of associated genes: 1324

No. of associated diseases: 31

click a disease name below to view its shared genes with current disease

Disease name	No. of shared genes
Systemic lupus erythematosus	467
Multiple sclerosis	439
Inflammatory bowel disease	407
Psoriasis	391
Crohn's disease	385
Ulcerative colitis	343
Systemic sclerosis	294
Atopic dermatitis	218
Sarcoidosis	188
Ankylosing spondylitis	173

Figure 3 An interactive network view showing the relationships among ADs

A network view on the relationships among ADs is shown on the left. Nodes represent different ADs, while edges represent their shared genes, with the width of the edges proportional to the number of genes shared by the two involved ADs. Users can move the nodes for better view by clicking and dragging the nodes. When the mouse hovers over a node, all the direct neighbors and connecting edges of the node are highlighted in yellow (1). Additional information, including the number of genes associated with the AD (2) and the list of ADs that they share associated genes with, is also shown. By clicking any AD in the list (3), users will be redirected to a new page with a list of shared genes between the two ADs. All the nodes and edges of the network are also clickable. Line widths are proportional to the numbers of sharing genes between diseases; edges are color coded to show the same information, with darker gray indicating fewer sharing genes and darker red indicating more sharing genes. See http://gaad.medgenius.info/diseases/ for details.

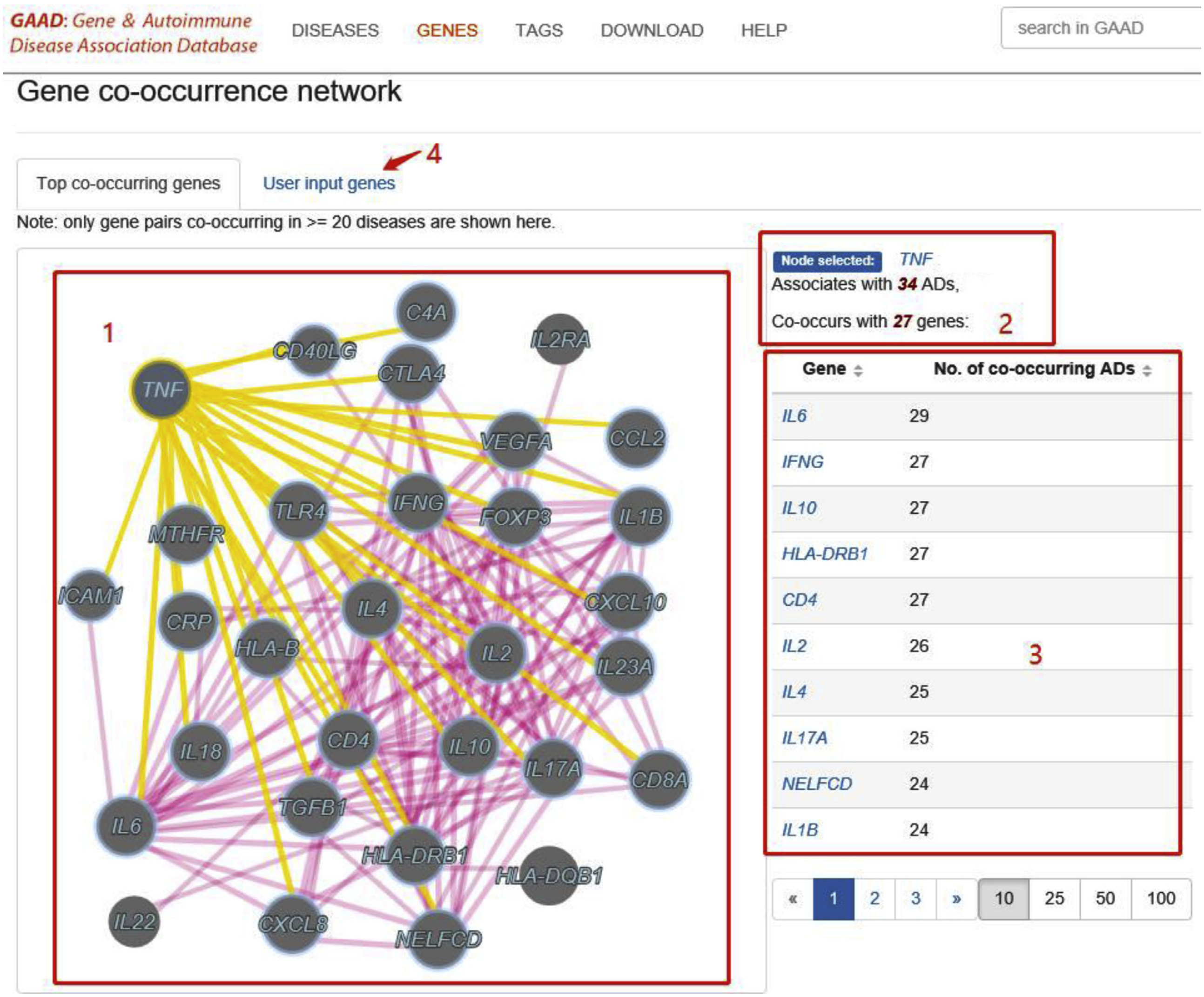
connects all our collected ADs (**Figure 3**, left panel); on average, an AD shares ~30% of its associated genes with other types of ADs.

To allow users to explore the relationships among ADs and their shared genes, at the bottom of the "DISEASES" page, we provide an interactive view for the relationships. As shown in Figure 3 (left panel), circles (nodes) represent ADs, with their diameters being proportional to the numbers of associated genes; the connecting lines (edges) represent their relationships, with the line width being proportional to the numbers of shared genes. Mouseover of a node highlights all its connecting edges and direct neighbors in the network (Figure 3, left panel). In addition, a list of its connected ADs is also shown (Figure 3, right panel). The nodes and edges are clickable; when clicked, the users will be redirected to new pages with detailed information on the AD and the shared genes between two ADs, respectively.

In addition, we also provide a list of significantly co-occurring gene pairs in GAAD. In total we have identified 5440 significantly co-occurring gene pairs, which can be found on the "GENES" page. At the bottom of the page, there is also an interactive visualization widget, allowing users to explore the gene co-occurrence network constructed using these gene pairs.

As shown below, immune response-related genes are tightly connected. For example, *TNF*, which encodes the tumor necrosis factor, significantly co-occur with other 27 out of in total 30 top co-occurring genes (gene pairs co-occurring in $\geq$20 diseases), most of which are also immune response related (**Figure 4**). These observations are consistent with previous findings that these genes often contribute to ADs independently and have additive effects on disease risk. In addition to this pre-defined network, users can construct their own network with user-supplied genes (Figure 4).

We queried the KEGG database to see whether the two genes of a pair tend to be involved in the same metabolic or signaling pathway, and/ or share functional similarities according to their annotations in the GO database. Compared to the randomly generated and non-significant gene pairs, significantly



Gene co-occurrence network

Note: only gene pairs co-occurring in >= 20 diseases are shown here.

Node selected: *TNF*
Associates with ***34*** ADs,
Co-occurs with **27** genes:

Gene	No. of co-occurring ADs
IL6	29
IFNG	27
IL10	27
HLA-DRB1	27
CD4	27
IL2	26
IL4	25
IL17A	25
NELFCD	24
IL1B	24

Figure 4 Gene co-occurrence network

Gene co-occurrence network constructed from significantly co-occuring gene pairs (see "co-occurring gene pairs" on the GENES page); only gene pairs co-occurring in more than 20 diseases are shown here. Essential elements of this part include: an interactive visualization of the network (1), detailed information of the highlighted node (2; *TNF* is highlighted in this case), a list of genes that co-occur with *TNF* in this network (3), and a widget allowing users to create their own co-occurrence network using genes of interest (4). See http://gaad.medgenius.info/genes/ for more details.

co-occurring pairs tend to have similar functions ($P < 2E{-}16$ for both KEGG and GO analyses, Fisher's exact test). In addition, we found that the more ADs in which both genes co-occur, the higher chance that the two genes are in the same pathway (**Figure 5**A) and share similar functions (Figure 5B).

Extraction of clinically-relevant information for associations

To make GAAD more straightforward to experimental biologists and medical scientists, we extracted clinically-relevant features describing the gene–disease associations from the text mining results, and assigned them into "tag groups". In total ten tag groups were created, each describing different aspects of the associations. These tag groups include *keyword of diagnostic value*, *onset stage*, *population* (in which the study was performed), *animal model*, *expression change* (between control and patient groups), *post-translational modification*, *associated microRNA*, *keyword* (used to do the query), *mutation type*, and *mutation position*. The "tag groups" are available at http://gaad.medgenius.info/tags/.

Example usage 1: Searching for putative causal genes for ADs

'Tags' are essentially keywords characterizing many aspects of the associations between genes and ADs, and are only available for text mining results. We grouped 'tags' into ten tag groups according to the characteristics of the corresponding associations, for example, population in which the association was described, the mutation type and position of the associated gene, and whether any potential diagnostic value of the association has been described. Information about the diagnostic value of an association would be most interesting

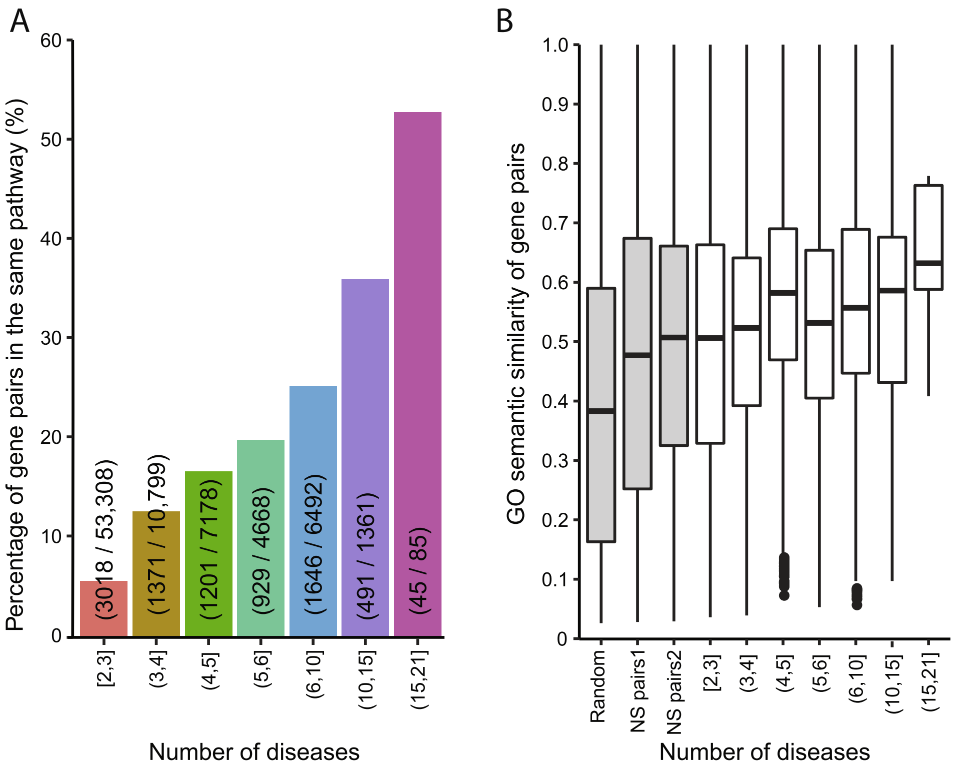
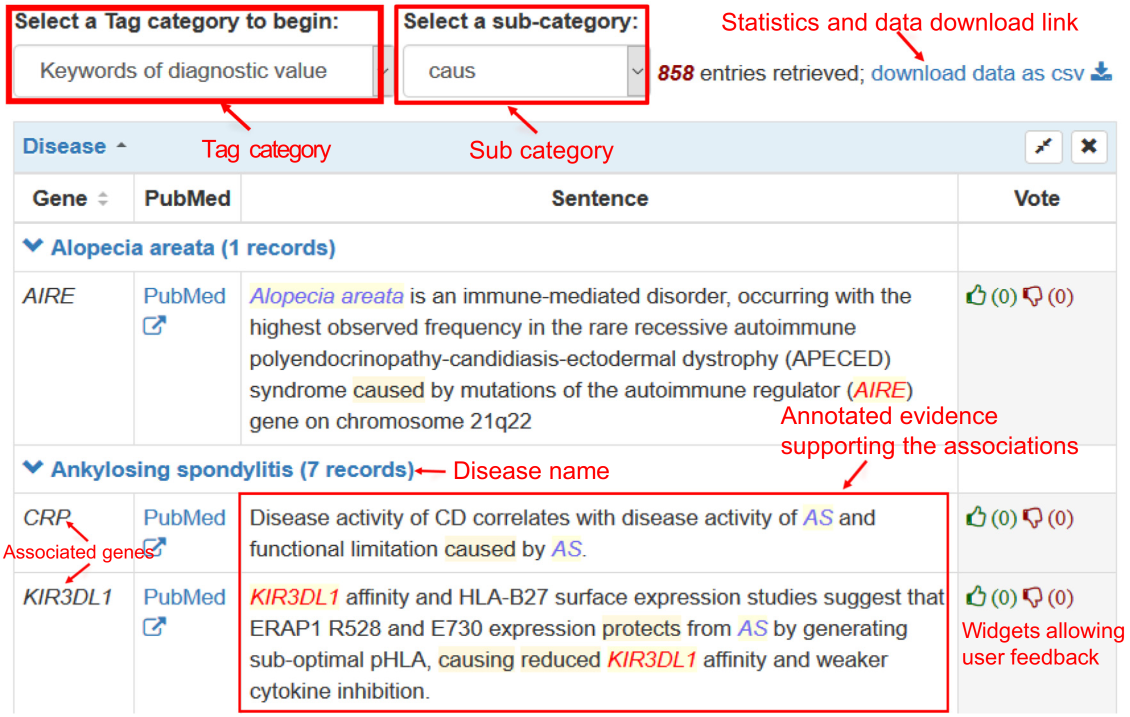


Figure 5 Genes in a co-occurring pair often have similar functions

The more ADs in which both genes of a pair co-occurring, the higher chance that they are in the same pathway (**A**) and functionally similar (**B**); see "Data analysis" for details on the calculation of functional similarities. Random refers to randomly generated pairs. NS pairs1 indicates gene pairs that do not co-occur in any ADs, whereas NS pairs2 indicates gene pairs co-occur in ADs but the co-occurrences are not significant ($P \geq 0.5$, Fisher's exact test; see "Data analysis" for details). The dots at the bottom of bins (4,5] and (6, 10] are outliners.


Statistics and data download link
Tag category
Sub category
Annotated evidence supporting the associations
Disease name
Associated genes
Widgets allowing user feedback

Select a Tag category to begin: Keywords of diagnostic value

Select a sub-category: caus

858 entries retrieved; download data as csv

Disease

Gene	PubMed	Sentence	Vote
Alopecia areata (1 records)			
AIRE	PubMed	*Alopecia areata* is an immune-mediated disorder, occurring with the highest observed frequency in the rare recessive autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED) syndrome caused by mutations of the autoimmune regulator (*AIRE*) gene on chromosome 21q22	(0) (0)
Ankylosing spondylitis (7 records)			
CRP	PubMed	Disease activity of CD correlates with disease activity of *AS* and functional limitation caused by *AS*.	(0) (0)
KIR3DL1	PubMed	*KIR3DL1* affinity and HLA-B27 surface expression studies suggest that ERAP1 R528 and E730 expression protects from *AS* by generating sub-optimal pHLA, causing reduced *KIR3DL1* affinity and weaker cytokine inhibition.	(0) (0)

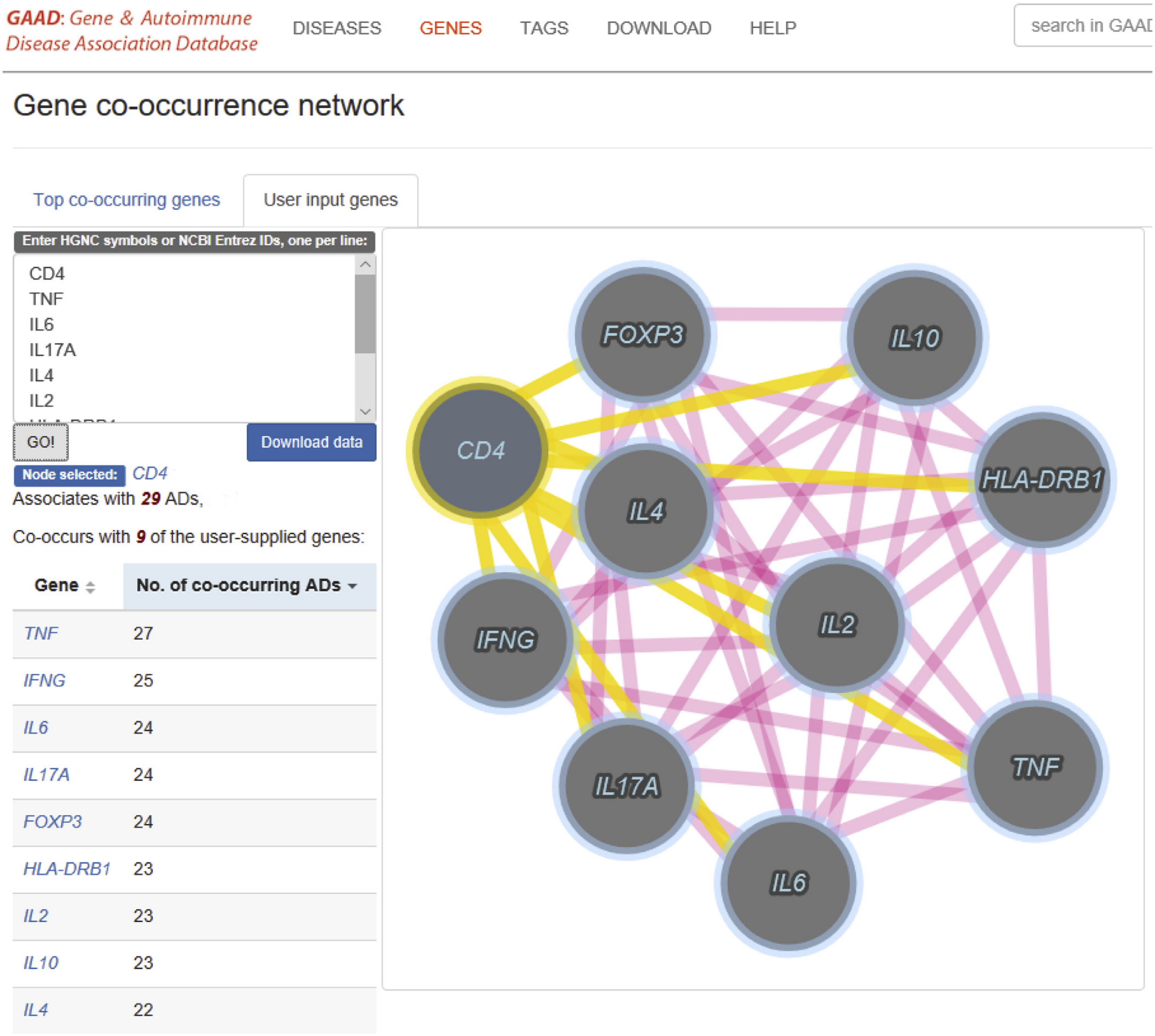
Figure 6 Searching for putative causal genes for ADs

Following the steps indicated on the screenshot, users are able to find putative causal genes for ADs collected in our database. The resulting entries (mostly are PubMed records) are organized according to the implicated diseases. Here "Tag category" refers to the major categories (*i.e.*, manually curated biologically meaningful categories) such as "Keyword of diagnostic value", "Mutation type" and "Onset stage", while "Sub category" refers to keywords belonging to the major category.

to clinicians. To find such information, users can simply go to the "TAGS" page, select the "Keywords of diagnostic value" from the "Tag category" dropdown menu. A second dropdown menu containing sub-categories is also available for users to further limit the number of matched entries (as shown in Figure 6). By default, the first sub-category from the second dropdown menu is always selected to avoid possible server overload due to the large number of matching entries under the corresponding "Tag category".

Example usage 2: analyzing gene co-occurrences in ADs with user-supplied genes

Gene pairs co-occurring in multiple ADs often share functional similarities and tend to be involved in the same metabolic or signaling pathways. On the "GENES" page we provide a list of pre-commutated co-occurring pairs and a network constructed with gene pairs co-occurring in more than 20 ADs. Using this module, users can also construct their own



Gene co-occurrence network

Enter HGNC symbols or NCBI Entrez IDs, one per line:

CD4
TNF
IL6
IL17A
IL4
IL2

Node selected: CD4

Associates with **29** ADs,

Co-occurs with **9** of the user-supplied genes:

Gene	No. of co-occurring ADs
TNF	27
IFNG	25
IL6	24
IL17A	24
FOXP3	24
HLA-DRB1	23
IL2	23
IL10	23
IL4	22

Figure 7 Analyzing gene co-occurrences in ADs with user-supplied genes
At the bottom of the "GENES" page, users can enter a list of genes of interest into the "User input genes" textbox to generate a gene co-occurrence network for these genes. Shown here is a network constructed from the top 10 genes that are associated with more than 25 ADs collected in our database. The 10 genes are *TNF*, *CD4*, *IL6*, *HLA-DRB1*, *IL2*, *IL10*, *IFNG*, *IL17A*, *IL4*, and *FOXP3*. Selected node ("*CD4*" in this case) will be outlined in yellow with its directly connected edges highlighted in yellow as well. Line widths are proportional to the numbers of co-occurring diseases between genes and edges are also color coded to show the same information, with darker gray indicating fewer co-occurring diseases and darker red indicating more co-occurring diseases.

gene co-occurrence network with a list of user-supplied genes. As shown in **Figure 7**, we submitted the ten genes that are associated with more than 25 ADs in GAAD and constructed a co-occurrence network. These genes are often interconnected within this network, suggesting that many of them are also likely to be associated with the same ADs.

Perspectives and concluding remarks

In this study we described GAAD, a disease–gene association database for autoimmune diseases. We assembled data from public databases and MEDLINE documents. We verified the associations by several rounds of manual inspection, especially for those derived from text mining to ensure the quality and credibility. To help users to better understand the interrelationships among ADs, we calculated shared genes among ADs and significantly co-occurring gene pairs. We provide powerful and easy-to-use web interface for users to access these data; we also include embedded tools so that users can query gene co-occurrences and construct customized co-occurrence network with genes of interest. In addition, associations were grouped according to relevant information, including the putative diagnostic value of the associations, type and position of gene polymorphisms, expression changes of implicated genes and the phenotypical consequences; such information should be of interest to experimental biologists and medical scientists. GAAD is freely available at http://gaad.medgenius.info.

Authors' contributions

GL, WC, and SM conceived the study, and calculated gene co-occurrences for all possible pairs; WC and XH designed the website. GL, XH, WC, and SM performed text mining and searched for gene–disease associations. XH drafted the manuscript; GL and WC edited the final manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

We would like to thank S. Balakrishnan, a member of our laboratory, for proofreading the manuscript.

References

[1] Davidson A, Diamond B. Autoimmune diseases. N Engl J Med 2001;345:340–50.

[2] Dayan CM, Daniels GH. Chronic autoimmune thyroiditis. N Engl J Med 1996;335:99–107.

[3] Salama AD, Levy JB, Lightstone L, Pusey CD. Goodpasture's disease. Lancet 2001;358:917–20.

[4] Lerner A, Jeremias P, Matthias T. The world incidence and prevalence of autoimmune diseases is increasing. Int J Celiac Dis 2015;3:151–5.

[5] Huang H, Fang M, Jostins L, Umicevic Mirkov M, Boucher G, Anderson CA, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature 2017;547:173–8.

[6] Karopka T, Fluck J, Mevissen HT, Glass A. The Autoimmune Disease Database: a dynamically compiled literature-derived database. BMC Bioinformatics 2006;7:325.

[7] Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 2012;491:119–24.

[8] Mu Q, Zhang H, Liao X, Lin K, Liu H, Edwards MR, et al. Control of lupus nephritis by changes of gut microbiota. Microbiome 2017;5:73.

[9] Zhang X, Zhang D, Jia H, Feng Q, Wang D, Liang D, et al. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat Med 2015;21:895–905.

[10] Maeda Y, Takeda K. Role of gut microbiota in rheumatoid arthritis. J Clin Med 2017;6:60.

[11] Shahi SK, Freedman SN, Mangalam AK. Gut microbiome in multiple sclerosis: the players involved and the roles they play. Gut Microbes 2017;8:607–15.

[12] Kohling HL, Plummer SF, Marchesi JR, Davidge KS, Ludgate M. The microbiota and autoimmunity: their role in thyroid autoimmune diseases. Clin Immunol 2017;183:63–74.

[13] Jin Y, Andersen G, Yorgov D, Ferrara TM, Ben S, Brownson KM, et al. Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants. Nat Genet 2016;48:1418–24.

[14] Boelaert K, Newby PR, Simmonds MJ, Holder RL, Carr-Smith JD, Heward JM, et al. Prevalence and relative risk of other autoimmune diseases in subjects with autoimmune thyroid disease. Am J Med 2010;123:183.e1–9.

[15] Henderson RD, Bain CJ, Pender MP. The occurrence of autoimmune diseases in patients with multiple sclerosis and their families. J Clin Neurosci 2000;7:434–7.

[16] Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. Curr Protoc Bioinformatics 2016;54:1.30.1–33.

[17] MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 2017;45:D896–901.

[18] Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics 2013;29:1830–1.

[19] Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics 2010;26:976–8.