T-HOD: Text-mined Hypertension, Obesity, Diabetes Candidate Gene Database

Johnny Chi-Yang Wu[1], Hong-Jie Dai[1,3], Richard Tzong-Han Tsai[4], Wen-Harn Pan[2], Wen-Lian Hsu[1, 3*]

[1]Institute of Information Science, Acdemia Sinica, Taipei, Taiwan, R.O.C.
[2]Institute of Biomedical Sciences, Acdemia Sinica, Taipei, Taiwan, R.O.C.
[3]Dep. of Computer Science, National Tsing-Hua Univ., HsinChu, Taiwan, R.O.C.
[4]Dept. of Computer Science & Engineering, Yuan Ze Univ., Taoyuan, Taiwan, R.O.C.

[*]Corresponding author: Tel: +886-2-2788-3799 ext. 1804, E-mail: hsu@iis.sinica.edu.tw

Abstract

Text mined hypertension, obesity, diabetes candidate gene database (T-HOD) is a database developed to collect lists of genes that are associated with three kinds of cardiovascular diseases – hypertension, obesity and diabetes, with the last disease specified into Type 1 and Type 2. T-HOD employed the state-of-art text-mining technologies, including a gene mention recognition/gene normalization system and a disease-gene relation extraction system, which can be used to affirm the association of genes with the three diseases and provide more evidence for further studies. The primary inputs of T-HOD are the three kinds of diseases, and the output is a list of disease-related genes which can be ranked based on their number of appearance, protein-protein interactions and single nucleotide polymorphisms. Currently, 837, 835, and 821 candidate genes are recorded in T-HOD for hypertension, obesity and diabetes, respectively. We believe T-HOD can help life scientists in search for more disease candidate genes in a less time- and effort-consuming manner. T-HOD is available at http://bws.iis.sinica.edu.tw/THOD.

Introduction

In biomedical literature mining, an important task is to identify disease related genes as its bio-signature. Previous work such as the genetic association database (GAD)[1], which is built on the basis of manual processes, has high manpower costs owing to the difficulty of maintaining these databases with the large volume of ever-growing new literatures. In addition, the study of disease pathogenesis has become increasingly difficult because of the diverse factors involved in disease progression. In most cases, development of these diseases is modulated by the variations of multiple genes and their interactions with environmental factors. Therefore, elucidating the pathogenic mechanisms of diseases turns out to be a demanding task. To investigate the complex genetics behind diseases systematically, it is necessary to integrate the finding of both small-scale studies and high-throughput researches. However, there are only a few databases and review papers that compile HOD related genes from the literature.

In the field of diabetes, T1Dbase [2] integrates valuable information on candidate genes from several databases for type 1 diabetes, while T2D-Db [3] compiles from PubMed human, mouse and rat genes involved in the pathogenesis of type 2 diabetes. For obesity genetics, the review paper “The Human Obesity Gene Map: The 2005 Update” [4] lists candidate genes and/or potential loci up until the end of 2005. For hypertension genes, GAD lists hundreds of hypertension candidate genes along with genes for several other diseases. All of the above mentioned resources were compiled manually. However, due to limited human resources, such databases cannot always be kept up-to-date. In recent years, various groups have proposed using automated text mining approaches to reduce human effort in constructing and updating such databases [5-9]. SNPs3D [8] and PubMeth [5] are two such databases constructed using text-mining approaches coupled with manual review and annotation steps. SNPs3D compiles candidate genes and single nucleotides polymorphism (SNP) sites related to cancers, neurodegenerative diseases and metabolic syndromes. PubMeth contains information on DNA methylation for several cancers. These two databases extract gene names that have a high co-occurrence with the target diseases. However, using the co-occurrence-based approach alone tends to yield a huge number of false-positive relations because of the lack of syntactic and semantic analysis.

Our database, T-HOD employed the state-of-art text-mining technologies we recently developed, including a gene mention recognition/normalization (GN) system [10-12] and a disease-gene relation extraction system [13]. Since gene names vary a great deal, different genes may contain the same name. Moreover, gene names may be ambiguous and easily confused with terms employed in other research fields. Our GN system was designed to alleviate the above problems, which was used to recognize gene terms and normalize them to their corresponding Entrez Gene IDs. In addition, we recently achieved promising results in extracting hypertension-related genes [13]. We extended and optimized the above systems to extract HOD genes in our T-HOD database.

T-HOD Interface and Implementation

As shown in Figure 1, the interface of T-HOD is divided into four regions. We will elucidate the function of each region in the following section, respectively.

Region 1: Control bar

Region 1 at the top of the frame contains a pull-down display menu. By clicking on the menu, users can select the disease of interest (Hypertension, Obesity, or Type 1/2 diabetes). Users can also decide whether to show specific gene information or use our advanced search function in this region.

Region 2: Candidate Gene list

After disease selection, Region 2 shows a list of curated candidate genes. Along each candidate gene, the list also displays the number of papers containing evidence sentences, as well as the number of SNPs and number of PPIs in separate columns. The list can be sorted by clicking on the column header, and it is accessible by hitting the “download” button at the bottom.

Region 3: Viewers

Region 3 provides several viewers, including sentence viewer, network viewer, advanced search option tabs, and statistics viewer. Users can switch between different viewers by clicking on the upper tags in this region.

Sentence Viewer: The sentence viewer provides curated evidence sentences for each selected candidate gene. If the candidate genes possess corresponding SNP information, the SNP evidence sentence would also be shown below the candidate gene evidence sentences. For each evidence sentence, the sentence viewer shows the source article’s PMID and year of publication with highlighted gene and disease terms. Display of the system can be adjusted by changing the font size of the texts. And in respect of valuable feedbacks, we constructed a user friendly interface for users to express their thoughts. In addition, for those who are interested in our database and plan to adopt its use in other studies, the information of T-HOD is attainable by hitting the “download” button below the gene list and supporting sentences, allowing them to acquire the disease-related genes and their supporting proof, respectively.

Network Viewer: Figure 2 shows the network viewer that presents a graphic-based gene-gene network for a selected candidate gene. For each selected candidate gene, the viewer integrates the corresponding PPI information recorded in the Human Protein Reference Database HPRD [14] to illustrate the gene-gene network. It allows users to discover the relations among extracted candidate genes. The blue node at the top of the window represents the gene that the user chose in Region 2. To cross examine the candidate genes, the user can double click on the nodes of other candidate genes shown in the same network. Accordingly, the network viewer will redraw the network graph based on the selected gene so that the user can navigate the database more smoothly.

Advanced Search: The advanced search option tab provides advanced search options that allow users to narrow down and specify the desired search results by the following items: publication date, Entrez Gene ID, gene name, and PubMed ID.

Statistics Viewer: The number of candidate genes and candidate SNP sites contained in T-HOD are summarized in the viewer. The statistics viewer also plots the number of candidate genes and the number of new candidate genes each year in bar charts as shown in Figure 3.

Region 4: Gene and SNP information

For each selected candidate gene, the information integrated from different resources is shown in Region 4. In this region, we integrate the following information from Entrez Gene and SNP database: the gene's official symbol, Entrez Gene ID, full name, synonyms and function summary. Users can also link to the corresponding database for further information.

Text mining-based Database Curation

Figure 4 shows the flowchart for constructing the T-HOD database. It is comprised of three stages: (1) Dataset Collection and Pre-processing, (2) Candidate Gene Extraction, and (3) Content Verification. We collected abstracts on HOD from PubMed, and used text mining techniques to extract HOD candidate genes and SNPs from them. The T-HOD curators verify the extracted list and curate the knowledge into the T-HOD database. In the following sub-sections, we will describe each stage in detail.

Stage 1: Dataset Collection and Preprocessing

In this stage, we collect HOD-related abstracts from PubMed and filter out those that are non-genetic. The filtered dataset are then pre-processed by several text mining components. After pre-processing, the genetic-related abstracts were split into sentences associated with section tags by using our section categorization component [15], such as "Results" and "Conclusion", which indicate their corresponding sections.

Stage 2: Candidate Genes Extraction

In Stage 2, we extract HOD-related candidate genes from the pre-processed dataset through the following steps. First, we employ a disease named entity recognition (NER) system to recognize disease terms in a sentence. Second, a GN system is used to recognize and normalize mentioned genes to their corresponding Entrez Gene IDs. Based on the results of the previous steps, if a disease term and a gene are present in the same sentence, they are extracted as a disease-gene (D-G) candidate pair. Finally, the D-G relation extraction system determines whether a relation indeed exists within this D-G pair.

Stage 3: Manual curation

While the employed text mining components have shown satisfactory scores (*cf.* Table 1), the text mined candidate genes are examined by all T-HOD curators in Stage 3 to further ensure the quality of the curated content. In this stage, newly extracted candidate genes and their corresponding evidence sentences and abstracts are presented to the T-HOD curators. T-HOD curators review each extracted candidate gene and remove the incorrect results. Currently, the curation process has only been done on abstracts before 2011. Because all annotated error cases are recorded to our SQL database, we can also use such data to modify our text mining components efficiently.

Proposed Task for Biocuration

For Track III–interactive text mining and user evaluation task of the BioCreative 2012 Workshop, we propose the following task for biocuration.

HOD Curation Task

When given a set of abstracts (compiled from those published in 2011) related to a specific disease, a bio-curator should:

1. Identify whether the abstracts contain disease-related gene information (curatable abstracts).
2. As for curatable abstracts, extract the following information: PMID of the abstract, gene terms and its corresponding gene ID from Entrez Gene, disease terms, relation assertion (positive or negative), and the evidence sentence containing the gene-disease pair.

Figure 5 shows the formal task descriptions provided to bio-curators. Note that since the abstract set used for the HOD curation task is publications from 2011, it is not yet verified by our T-HOD curators.

The bio-curator then compares their manually curated results with the text-mined results processed by T-HOD. For the convenience of bio-curators in analyzing the results, we developed an interface that directly provides the information of T-HOD in the desired output format with additional PubMed and Entrez Gene links. These results are also available for download. Furthermore, this interface is able of notifying the curators when an abstract is not found in our database, or it does not contain any relations of interest. An example of the interface output is shown in Figure 6. This interface is available at http://bws.iis.sinica.edu.tw:8080/THOD/request_sentence_list.

Results and Discussion

Evaluation of a candidate gene database is difficult. Different standards and perspectives can produce different results. In our previous work [13], the employed D-G extraction system has shown satisfactory area-under-curve (AUC) scores of 81.4% and 83% for hypertension and diabetes, respectively. In this work, we compared the performance of T-HOD with the contents of GAD. The bench marking results are shown in Table 1. Disease-related literatures that exist in both databases were chosen for evaluation. Performance for the identification of gene-disease relations in hypertension, obesity and type 2 diabetes documents all achieved a score around 75%. In contrast, relations of type 1 diabetes only obtained a score around 70%.

There are several possible reasons that may result in the difference between T-HOD and GAD results. In order for curating a gene-disease relation into our T-HOD, the identity of both the candidate gene and disease terms must be normalized. In the current implementation, gene terms are normalized by a collective entity disambiguation method [16] to its corresponding Entrez Gene ID, while disease terms are recognized through a list of vocabularies. Error in the normalization of genes and the imperfect list of disease terms utilized may lead to the loss of relations that are present within documents. In addition, the difficulty of extract D-G relation will increase when one or both the disease and gene are expressed with an anaphoric expression. Furthermore, T-HOD only recognizes D-G relations within the same sentence. Cross sentence relations are currently not available, but it is a topic worth studying in the future. Cooper and

Kershenbaum [17] has identified co-reference as one of the reasons for the decreasing recall in biomedical relation extraction task. Finally, determining negations within a sentence is also an important issue. The present T-HOD system can only deal with negation descriptions in basic phrase structures, which may be insufficient in distinguishing more complex negation narratives.

Funding

This work was supported by the National Core Facility Program for Biotechnology, Taiwan (Bioinformatics Consortium of Taiwan, NSC100-2319-B-010-002).

References

1. Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database**. *Nature Genetics* 2004, **36**(5):431-432.
2. Hulbert EM, Smink LJ, Adlem EC, Allen JE, Burdick DB, Burren OS, Cassen VM, Cavnor CC, Dolman GE, Flamez D *et al*: **T1DBase: integration and presentation of complex data for type 1 diabetes research**. *Nucleic Acids Res* 2007, **35**(Database issue):D742-746.
3. Agrawal S, Dimitrova N, Nathan P, Udayakumar K, Lakshmi SS, Sriram S, Manjusha N, Sengupta U: **T2D-Db: an integrated platform to study the molecular basis of Type 2 diabetes**. *BMC Genomics* 2008, **9**:320.
4. Rankinen T, Zuberi A, Chagnon YC, Weisnagel SJ, Argyropoulos G, Walts B, Pérusse L, Bouchard C: **The Human Obesity Gene Map: The 2005 Update**. *Obseity* 2006, **14**:529-644.
5. Ongenaert M, Van Neste L, De Meyer T, Menschaert G, Bekaert S, Van Criekinge W: **PubMeth: a cancer methylation database combining text-mining and expert annotation**. *Nucleic Acids Res* 2008, **36**(Database issue):D842-846.
6. Hahn U, Wermter J, Blasczyk R, Horn PA: **Text mining: powering the database revolution**. *Nature* 2007, **448**(7150):130.
7. Fang YC, Huang HC, Chen HH, Juan HF: **TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining**. *BMC Complement Altern Med* 2008, **8**:58.
8. Yue P, Melamud E, Moult J: **SNPs3D: Candidate gene and SNP selection for association studies**. *BMC Bioinformatics* 2006, **7**:-.
9. Fang YC, Lai PT, Dai HJ, Hsu WL: **MeInfoText 2.0: gene methylation and cancer relation extraction from biomedical literature**. *BMC Bioinformatics* 2011, **12**(1):471.
10. Dai H-J, Lai P-T, Tsai RT-H: **Multistage Gene Normalization and SVM-Based Ranking for Protein Interactor Extraction in Full-Text Articles**. *IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS* 2010, **7**(3):412-420.
11. Tsai RT-H, Sung C-L, Dai H-J, Hung H-C, Sung T-Y, Hsu W-L: **NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition**. *BMC Bioinformatics* 2006, **7**(Suppl 5):S11.
12. Dai H-J, Chang Y-C, Tsai RT-H, Hsu W-L: **Integration of gene normalization stages and co-reference resolution using a Markov logic network**. *Bioinformatics* 2011, **27**(18):2586-2594.
13. Tsai RT-H, Lai P-T, Dai H-J, Huang C-H, Bow Y-Y, Chang Y-C, Pan W-H, Hsu W-L: **HypertenGene: Extracting key hypertension genes from biomedical literature with position and automatically-generated template features**. *BMC Bioinformatics* 2009, **10**(Suppl 15):S9.
14. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A *et al*: **Human Protein Reference Database--2009 update**. *Nucleic Acids Res* 2009, **37**(Database issue):D767-772.

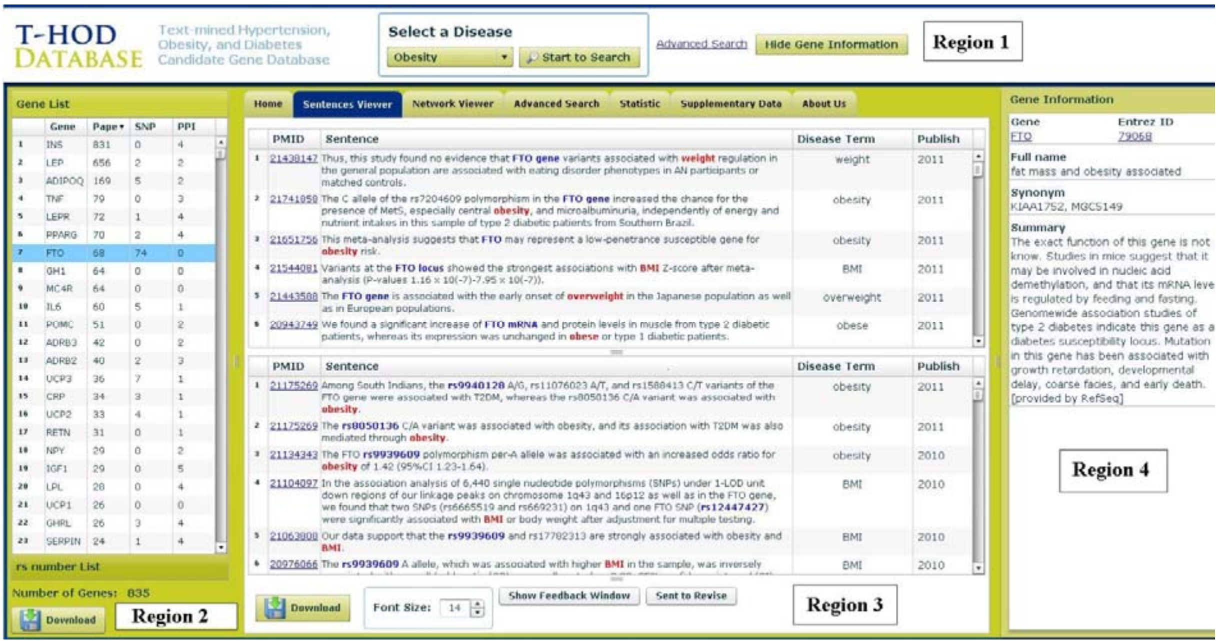
15. Lin RTK, Dai H-J, Bow Y-Y, Chiu JL-T, Tsai RT-H: **Using conditional random fields for result identification in biomedical abstracts** *Integrated Computer-Aided Engineering* 2009, **16**(4):339-352.

16. Dai H-J, Tsai RT-H, Hsu* W-L: **Entity Disambiguation Using a Markov-Logic Network**. In: *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP); Chiang Mai, Thailand*. 2011: 846-855.

17. Cooper JW, Kershenbaum A: **Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information**. *BMC Bioinformatics* 2005, **6**(1):143.

Table 1. Comparison of candidates genes in T-HOD with GAD.

	True Positive	**False Positive**	**False Negative**	**Precision**	**Recall**	**F score**	**Number of Documents**
Hypertension	165	42	49	0.797	0.771	0.784	150
Obesity	105	35	29	0.75	0.784	0.766	115
Type 1 Diabetes	60	24	27	0.714	0.69	0.702	73
Type 2 Diabetes	127	35	37	0.784	0.774	0.779	140
Overall	457	136	142	0.771	0.763	0.767	608



	Gene	Pape▾	SNP	PPI
1	INS	831	0	4
2	LEP	656	2	2
3	ADIPOQ	169	5	2
4	TNF	79	0	3
5	LEPR	72	1	4
6	PPARG	70	2	4
7	FTO	68	74	0
8	GH1	64	0	0
9	MC4R	64	0	0
10	IL6	60	5	1
11	POMC	51	0	2
12	ADRB3	42	0	2
13	ADRB2	40	2	3
14	UCP3	36	7	1
15	CRP	34	3	1
16	UCP2	33	4	1
17	RETN	31	0	1
18	NPY	29	0	2
19	IGF1	29	0	5
20	LPL	28	0	4
21	UCP1	26	0	0
22	GHRL	26	3	4
23	SERPIN	24	1	4

Number of Genes: 835

	PMID	Sentence	Disease Term	Publish
1	21430147	Thus, this study found no evidence that **FTO gene** variants associated with **weight** regulation in the general population are associated with eating disorder phenotypes in AN participants or matched controls.	weight	2011
2	21741058	The C allele of the rs7204609 polymorphism in the **FTO gene** increased the chance for the presence of MetS, especially central **obesity**, and microalbuminuria, independently of energy and nutrient intakes in this sample of type 2 diabetic patients from Southern Brazil.	obesity	2011
3	21651756	This meta-analysis suggests that **FTO** may represent a low-penetrance susceptible gene for **obesity** risk.	obesity	2011
4	21544081	Variants at the **FTO locus** showed the strongest associations with **BMI** Z-score after meta-analysis (P-values 1.16 × 10(-7)-7.95 × 10(-7)).	BMI	2011
5	21443588	The **FTO gene** is associated with the early onset of **overweight** in the Japanese population as well as in European populations.	overweight	2011
6	20943749	We found a significant increase of **FTO mRNA** and protein levels in muscle from type 2 diabetic patients, whereas its expression was unchanged in **obese** or type 1 diabetic patients.	obese	2011

	PMID	Sentence	Disease Term	Publish
1	21175269	Among South Indians, the **rs9940128** A/G, rs11076023 A/T, and rs1588413 C/T variants of the FTO gene were associated with T2DM, whereas the rs8050136 C/A variant was associated with **obesity**.	obesity	2011
2	21175269	The **rs8050136** C/A variant was associated with obesity, and its association with T2DM was also mediated through **obesity**.	obesity	2011
3	21134343	The FTO **rs9939609** polymorphism per-A allele was associated with an increased odds ratio for **obesity** of 1.42 (95%CI 1.23-1.64).	obesity	2010
4	21104097	In the association analysis of 6,440 single nucleotide polymorphisms (SNPs) under 1-LOD unit down regions of our linkage peaks on chromosome 1q43 and 16p12 as well as in the FTO gene, we found that two SNPs (rs6665519 and rs669231) on 1q43 and one FTO SNP (**rs12447427**) were significantly associated with **BMI** or body weight after adjustment for multiple testing.	BMI	2010
5	21063800	Our data support that the **rs9939609** and rs17782313 are strongly associated with obesity and **BMI**.	BMI	2010
6	20976066	The **rs9939609** A allele, which was associated with higher **BMI** in the sample, was inversely	BMI	2010

Font Size: 14

Gene	Entrez ID
FTO	79068

Full name
fat mass and obesity associated

Synonym
KIAA1752, MGC5149

Summary
The exact function of this gene is not know. Studies in mice suggest that it may be involved in nucleic acid demethylation, and that its mRNA leve is regulated by feeding and fasting. Genomewide association studies of type 2 diabetes indicate this gene as a diabetes susceptibility locus. Mutation in this gene has been associated with growth retardation, developmental delay, coarse facies, and early death. [provided by RefSeq]

Figure 1. User interface of the T-HOD database. The user interface is divided into four regions for precise introduction.

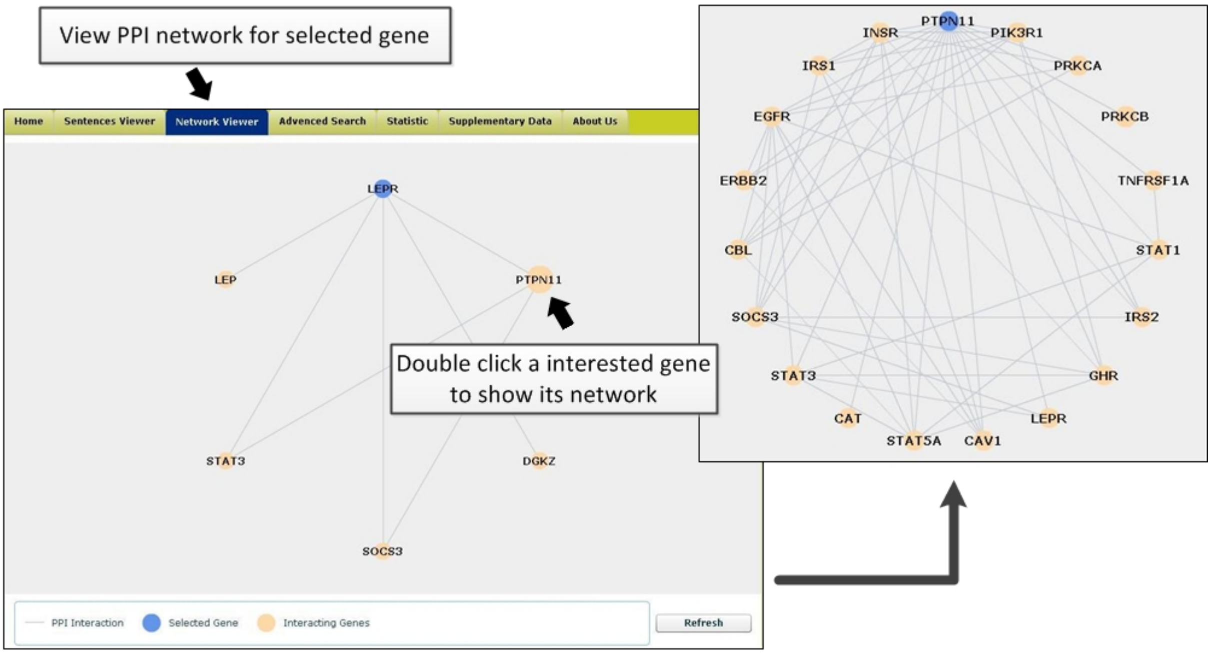


Figure 2. The network viewer of the T-HOD database.

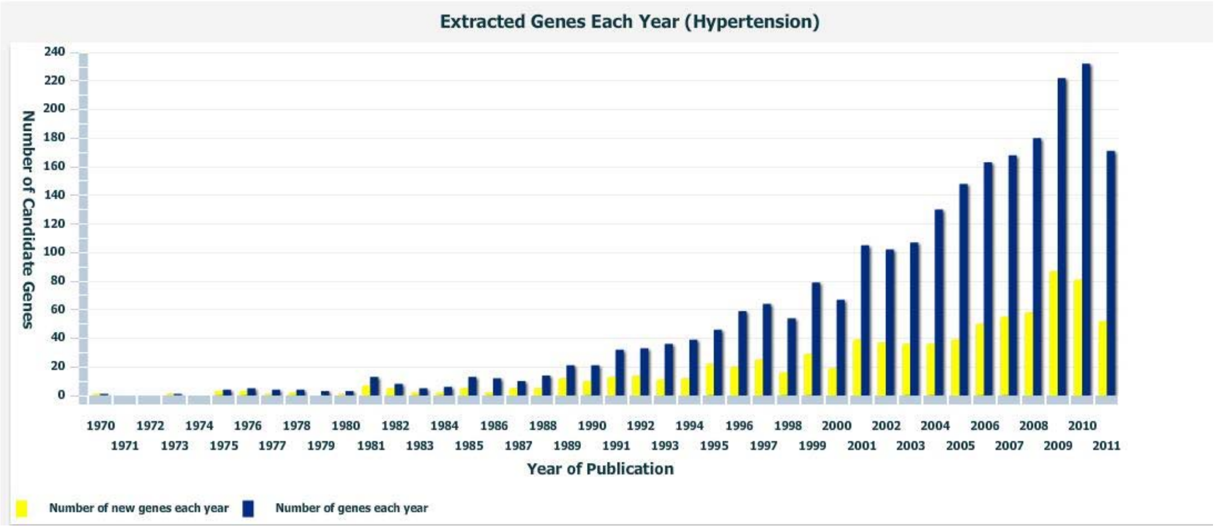


Figure 3. Statistics of the extracted hypertension candidate genes. The blue bars indicate the number of genes extracted each year, while the yellow bars specify the number of novel genes discovered each year.

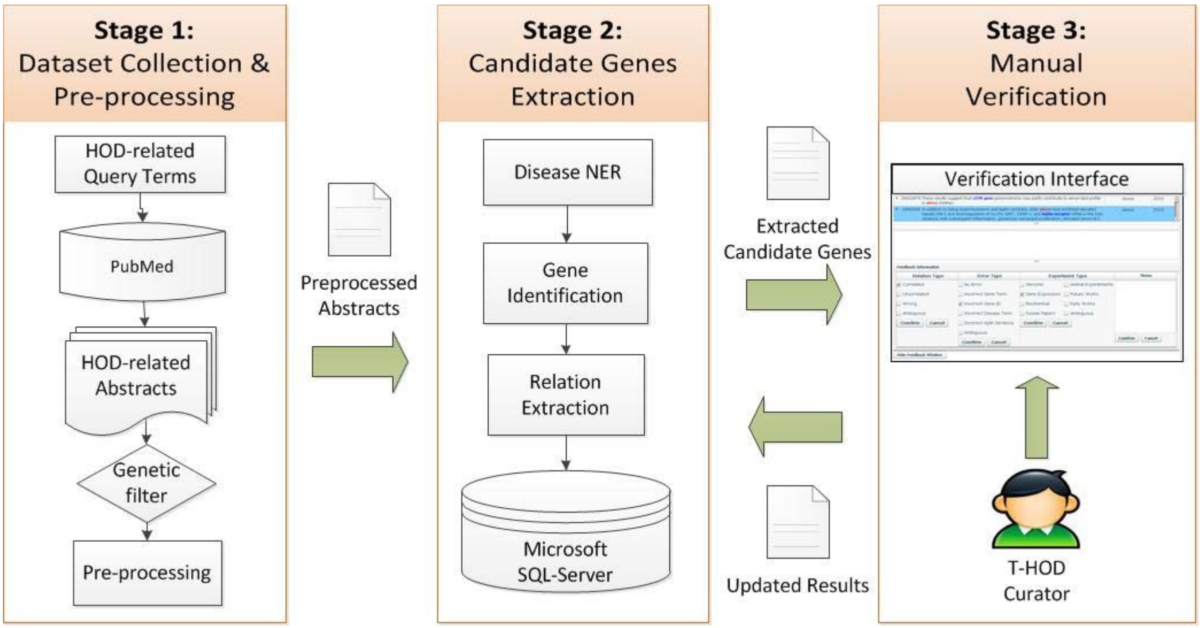


Figure 4. The flowchart of T-HOD database construction.

Manual Task: Curators will be given a list of PubMed abstracts for further processing, and should provide an output spreadsheet that contains the information of interest.

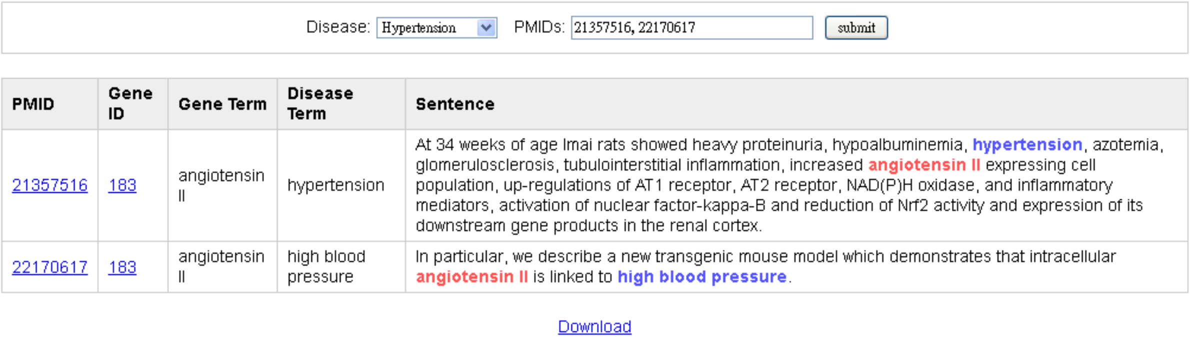
Using T-HOD: Curators will compare the information retrieved by T-HOD regarding the given set of abstracts with those that are extracted manually, analyze their differences and offer any suggestions for further improvement.

Input: Assigned set of specific disease-related abstracts.

Output: Output of the extracted information should be presented accordingly to the following format:

PMID | Gene ID | Gene Term | Disease term | Evidence sentence

Figure 5. Illustration of the proposed task for bio-curators.



Disease: Hypertension PMIDs: 21357516, 22170617

PMID	Gene ID	Gene Term	Disease Term	Sentence
21357516	183	angiotensin II	hypertension	At 34 weeks of age Imai rats showed heavy proteinuria, hypoalbuminemia, **hypertension**, azotemia, glomerulosclerosis, tubulointerstitial inflammation, increased **angiotensin II** expressing cell population, up-regulations of AT1 receptor, AT2 receptor, NAD(P)H oxidase, and inflammatory mediators, activation of nuclear factor-kappa-B and reduction of Nrf2 activity and expression of its downstream gene products in the renal cortex.
22170617	183	angiotensin II	high blood pressure	In particular, we describe a new transgenic mouse model which demonstrates that intracellular **angiotensin II** is linked to **high blood pressure**.

Figure 6. The interface for bio-curators.