A literature search tool for intelligent extraction of disease-associated genes

Jae-Yoon Jung,1 Todd F DeLuca,1 Tristan H Nelson,2 Dennis P Wall1,2

ABSTRACT
Objective To extract disorder-associated genes from the scientific literature in PubMed with greater sensitivity for literature-based support than existing methods.
Methods We developed a PubMed query to retrieve disorder-related, original research articles. Then we applied a rule-based text-mining algorithm with keyword matching to extract target disorders, genes with significant results, and the type of study described by the article.
Results We compared our resulting candidate disorder genes and supporting references with existing databases. We demonstrated that our candidate gene set covers nearly all genes in manually curated databases, and that the references supporting the disorder–gene link are more extensive and accurate than other general purpose gene-to-disorder association databases.
Conclusions We implemented a novel publication search tool to find target articles, specifically focused on links between disorders and genotypes. Through comparison against gold-standard manually updated gene–disorder databases and comparison with automated databases of similar functionality we show that our tool can search through the entirety of PubMed to extract the main gene findings for human diseases rapidly and accurately.

BACKGROUND
With the advance of genotyping and sequencing technologies, a rising number of studies have reported genetic association with various disorders in the past decade. As hundreds of genes may be involved in one complex disorder, a thorough literature review is fundamental to understanding genetic risk factors of any given human disorder. For example, if we are interested in the genetic etiology of schizophrenia, we would first like to know which genes have been reported for association with the most important literature evidence to justify the association. However, few applications have been available to help search and keep track of up-to-date, gene–disease associations. Many sites provide detailed gene data, including GeneCards,1 PharmGKB,2 WikiGenes,3 and iHOP4 but it is not easy to find the disorders associated with a given gene using these sites. Another group of tools, including LitInspector,5 MuGeX,6 Querite (http://querite.info) and NEXTBIO genetic markers (http://nextbio.com) provide supporting references and snippets of text from abstracts highlighting the target gene or disorder, but their genetic and/or disease coverage is limited and the methods, for gene–disorder association (such as simple co-occurrence of terms), often yield high rates of false positives. Disorder-oriented sites like SFARI gene7 and SZGene8 provide candidate genes for a target disorder with or without references, but cover only a single disorder. Previously we built a meta-search tool that integrates results from several of these sites.9 Using this tool, we discovered significant discrepancies between databases, with few providing adequate references to supporting literature.

Provided this context, our goal in the present study was to build a novel PubMed extraction tool that focuses on identifying target disorders and associated human genes from all original research articles, and to compare the results from this tool with the existing databases that provide disorder candidate genes and supporting references, including Online Mendelian Inheritance in Man (OMIM),10 HuGE Navigator,11 and Genetic Association Database (GAD).12 Various text-mining algorithms have been proposed to address entity recognition in scientific literature13 14 and infer novel gene–disorder relationships.15–20 We set our scope in this work, however, precisely to extract gene–disease linkages reported in the research articles, rather than to discover new associations based on literature information.

METHODS
Retrieving disorder-related articles
For each target disorder name, we built a comprehensive PubMed query to retrieve disorder-specific research or review articles. First, we mapped the given disorder name to the corresponding medical subject headings (MeSH),21 and obtained target disorder aliases from MeSH, MedlinePlus,22 and Genetics Home Reference.23 After expanding names for plurals (eg, disorder → disorders) and synonyms (eg, syndrome ↔ disorder/disease), we required in the query that these names and aliases appear in the title or MeSH entries sections of papers. We did not search in any field or in abstracts as we observed false-positive findings (ie, articles that are not directly related to the target disorder) when we allowed these fields. Second, we filtered out articles with publication types that were not relevant to research or reviews, for example bibliographies, comments, and editorials. We also limited results to publication dates after 1990, as we were interested in retrieving recent, genetics-oriented research articles. This was a conservative publication date filter, considering that the human genome project began in 1990 and the pilot phase of sequencing was done in 1999.24 An example of the expanded query targeted to autism spectrum disorder (ASD) is shown in figure 1. We used E-utils25 to execute this query, extracting PubMed article identifiers and details for further steps.
Screening genetics-related articles

In the first step, we retrieved articles related to the target disorder; we then performed further steps to reduce this set to include only papers related to genetics. First, if an article has MeSH annotations, we examined whether they include terms under ‘genetic techniques’ sub-tree; terms under ‘genetic phenomena’ sub-tree; or ‘genetics’ subheading. Second, if the article has no MeSH annotation, we examined if the title or abstract include genetics-related keywords. We obtained these keywords by comparing two training sets of abstracts. In particular, we selected 32 disorders from the genome-wide association studies (GWAS) catalog and downloaded two sets of abstracts from PubMed: those that include MeSH terms fitting in above conditions (158 745 articles); and those that have MeSH terms, but do not include terms of the first set (385 383 articles). After removing common frequent words based on Corpus of Contemporary American English, we measured the word frequency and compared the top 5000 words from each set in order to find out keywords that uniquely or dominantly (ie, top 1% after Wilcoxon signed-rank test) appear in the first set only. Supplementary table 1 (available online only) shows the top 20 keywords sorted by word frequency. We also used this keyword extraction method to identify abstract structures and study types that we explain below.

Analyzing structure of abstracts, study types and negations

As described in the background section, many existing tools still use simple co-occurrence to show gene–disorder associations. This is not reliable when we want to learn the exact study and reference in which specific associations are tested and reported. For example, sentences like ‘neuroligin genes have been associated with autism’ can occur in the introduction section of an abstract, but the main topic of the paper may not be relevant to neuroligin genes at all. Therefore we decided to use the structure of abstract to address this issue. We assume that the main findings of a research article must be reported in the result/conclusion sections of the abstract, or in the title, and in these locations only. This assumption enables us to separate tested genes in the background or methods section (eg, ‘We tested A, B, and C genes’) from associated genes (eg, ‘Only C gene was highly expressed’), and introductory statements in the background section (eg, ‘We previously showed that gene A is associated’). For abstracts without designated structure, we built a set of keywords and rules to identify them, using the available structured abstracts as training data. To extract the study type, we used publication types (eg, ‘reviews’ or ‘case report’) and MeSH terms (eg, ‘disease models, animal’ or ‘genome-wide association study’) when such information is available, or used keywords if MeSH terms or publication types are not annotated. We derived another set of rules to find negated statements in either the title or abstracts; for this we used example sentences obtained from BioNOT.

Gene representation

Finding gene symbols and their names in the literature and mapping them to unique identifiers is one of the major topics in biological literature mining, and a large number of algorithms exist to address this issue within various contexts. While following up our previous study, we recognized that many genetic–disorder-related articles only use gene symbols or protein symbols, rather than using full names. We tested two of the widely used entity recognition tools trained for human genes (ABNER and BANNER), but they did not show high precision for this type of task. To address this, we implemented a precision-based gene recognition procedure, which is similar to the protein name extraction algorithm of Fukuda et al or LitInspector. For each article, we first scanned the title and abstract to identify symbols that match up with gene patterns. For example, official human gene symbols can be identified with these regular expressions: /\[A-Z][A-Z0-9-]+/ or /\[\w\][0-9]+ or\[0-9]+/. When such a symbol was found, we determined the semantics of the symbol based on the context in which the symbol is located. For example, symbol ‘CGH’ may be used as an alias term of HTC2 gene, but may also mean array-CGH. We checked whether the immediate previous/following text around this symbol includes (1) in a list of gene symbols (eg, ‘X chromosome genes like DMD, MAOA, CGH, and FMR1’); (2) full (official) gene name (eg, ‘hypertrichosis 2 (CGH)’); (3) genetic keywords defined in the previous section (eg, ‘CGH deletion’); (4) other full names for the same pattern (eg, ‘comparative genome hybridization (CGH)’); or (5) other (dis)qualified for the same pattern (eg, ‘array CGH’). We kept the track of found symbols per article, assuming that a symbol can only have one meaning in the same article. The scan was performed twice, because the meaning of a symbol may not be decided in the current position, but in the later part of the title or abstract. The comparison output of our algorithm with ABNER and BANNER, including test input sentences and tagged words, is shown in the supplementary files (available online only).

Assessing the significance of articles and genes

Ranking of articles and terms based on the strength of publication and the structure of the article has been thoroughly studied by Demner-Fushman and Lin. We combined the temporal...
significant of articles with positive/negative findings of genes in order to assess the significance of an association between a gene and a target disorder. We defined the significance of an article based on the impact factor of the published journal and the publication year. The score of an article \(a\) for a given disorder \(d\) was defined as follows:

\[
\text{Score}(d, a) = (\text{Importance of the article}) \times (\text{Decay factor}) \\
= (1 + \log_2(\text{Impact Factor}(a) + 1)) \\
\times (1 - \lambda \times (\text{this year} - \text{Publication Year}(a)))
\]

Where \(\lambda = 1/((\text{this year} - 1990))\)

For example, if an article is published this year in a journal without a known impact factor, the score for the article is 1.0, and will decrease every year after publication. We used the decay factor to put priority on more recent findings. All other things being equal, recent articles will have slightly higher scores. The significance of a gene \(g\) for a given disorder \(d\) was defined by the sum of scores of articles reporting an association of gene \(g\) and disorder \(d\).

\[
\text{Score}(d, g) = \sum_a [\text{Score}(d, g) \\
\times \text{Association}(d, a, g) \times \text{AdjustStudyType}(a)]
\]

Where Association\((d, a, g)\) is (1) 0, if \(g\) is not one of the main findings of this article; (2) 1.0, if \(g\) is reported in the title/results/conclusion; or (3) –1.0 if \(g\) is one of the main findings, but the statement is negative. AdjustStudyType\((a)\) is defined as (1) 0, if the article is a review or hypothesis; (2) 0.5, if the article is a case study or examines blood/serum protein levels. The collective gene score can have a negative value when there is a preponderance of evidence against the association according to our scoring algorithm. The overall procedure of our search tool is summarized in figure 2.

**RESULTS**

**Reference coverage comparison with existing databases**

We tested our implementation with 10 complex disorders and genetic syndromes selected from the GWAS Catalog (attention deficit with hyperactivity disorder, ankylosing spondylitis, ASD, bipolar disorder, multiple sclerosis, and schizophrenia) and genetics home reference (Angelman syndrome, Down syndrome, Huntington disorder, and Lynch syndrome). Table 1 shows the number of gene–disorder association references we found and the number of such references from the union of HuGE Navigator, OMIM, and GAD. For each target disorder, our tool covered more references than the union of results from these sites and we confirmed that all of these articles are specific to the target disorder by manual inspection. We examined all the articles our tool did not retrieve, and found that a majority of them was not related to the target disorder. For example, the target disorder name may be stated in the abstract, but the main topic is for a different disorder. Other causes for exclusion included: the PubMed ID was not available for the article, the article was a commentary article, or the article was published before 1990. We also show high-profile reference samples that were not included in the compared repositories in table 1 and supplementary table 1 (available online only).

**Tested, positive result, or negative result genes**

As shown in the Methods section, we separately identified tested (or simply mentioned) genes, gene with positive findings, and genes with negative findings. By analyzing negating expressions, gene symbols, and disorder names that appear within the same sentence of the title, result, or conclusion sections, we found that a significant number of articles report negative associations or null findings, in which targeted genes showed no difference in case–control experiments, or genetic variants were not found in patient groups. Currently GAD is the only external resource that provides such references for multiple disorders, so we compared our result with those of GAD. As shown in table 2, our result covers more references than GAD for all target disorders. We examined articles shown only in GAD and found that some of the tested genes without positive association findings were reported as negative associations, while we only count genes combined with explicit negating statements in either title or abstract. We also show reference samples that are not included in GAD in table 2 and supplementary table 2 (available online only).
An example of a single disorder result: ASD

Identifying association study articles and categories

Next, we report a detailed result of single disorder case, ASD. As of October 2012, we identified 23,661 ASD-related articles in PubMed by querying disorder name and aliases plus MeSH terms without applying any filter. Our expanded query with filters, shown in figure 1, returned about half of them (12,984) as ASD-specific original research or review articles. Second, we identified 1,581 genetics-related publications among them by scanning them for targeted MeSH terms or keywords. Third, we categorized these articles in order to weight the scores by their publication or study types. We found 232 reviews (will not be included in scoring) and 135 case reports (50% of the original score will be applied) by publication types. We also found 37 animal models, 272 common variant-focused articles including GWAS, 146 rare variant-focused publications including copy-number variation (CNV) and exome sequencing studies, 52 gene expression-related articles, and 23 blood/serum protein level articles (50% of the original score will be applied) by publication types. In table 3 and supplementary table 3 (available online only), we summarize our categorization result with references, in order to demonstrate that our tool can effectively retrieve high-impact research articles related to a target disorder in different study types.

Assessment of candidate genes and references

We identified 597 gene symbols; 437 of them have their collective score greater than 1.0. Although the fragile X (FMR1) and methyl CpG binding (MECP2) genes have the largest number of associated articles, our result shows that CNTNAP2 (contactin-associated protein-like 2) is the highest score gene with more recent, high-profile publications. Table 4 and supplementary table 4 (available online only) show our top 10 candidate genes and supporting reference examples. To examine whether we found proper articles without missing a significant portion, we selected two external resources to compare our result with. First, HuGE Navigator maintains genetics-related publications using an algorithmic search, and we obtained 256 articles with a disease term of ‘autistic disorder’. Second, SFARI gene is a manually curated, ASD-specific database, and we obtained 278 articles that are associated with category 1 (confident) to 5 (minimal evidence) genes, according to their classification method. Compared with the set of genes from HuGE Navigator, our reference set missed one article primarily describing schizophrenia73 and not ASD. Similar cases were found for SFARI set; our result missed 48 articles; however, the main topic of such articles is not ASD specific but on comorbid disorders including epilepsy,86–87 intellectual disability,86–87 and attention-deficit hyperactivity disorder.88–89

Next, we compared our ASD candidate gene sets with those from external databases. Our set of 597 genes included (1) all candidate genes of GeneCards (31 genes) and PharmGkb (four genes); (2) 121/133 genes of category 1 to 4 in SFARI gene; (3) 21/22 syndromic genes in SFARI; and (4) 231/426 genes in HuGE Navigator. For all genes we missed in the SFARI set, gene names or symbols were not actually listed in the title or abstract. While we missed about a half of candidate genes in the

Table 2 References with negative results from our tool and GAD

<table>
<thead>
<tr>
<th>Target disorder</th>
<th>#Ref. with negative result</th>
<th>#Ref. with negative result in GAD</th>
<th>#Ref. missed in our result</th>
<th>Example ref. in our result only*</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>135</td>
<td>15</td>
<td>1</td>
<td>55</td>
</tr>
<tr>
<td>Angelman syndrome</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>56</td>
</tr>
<tr>
<td>Ankylosing spondylitis</td>
<td>114</td>
<td>13</td>
<td>0</td>
<td>57</td>
</tr>
<tr>
<td>ASD</td>
<td>168</td>
<td>31</td>
<td>3</td>
<td>58</td>
</tr>
<tr>
<td>Bipolar disorder</td>
<td>341</td>
<td>75</td>
<td>6</td>
<td>59</td>
</tr>
<tr>
<td>Down syndrome</td>
<td>185</td>
<td>3</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>Huntington disorder</td>
<td>82</td>
<td>3</td>
<td>2</td>
<td>61</td>
</tr>
<tr>
<td>Lynch syndrome</td>
<td>203</td>
<td>0</td>
<td>0</td>
<td>62</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>496</td>
<td>73</td>
<td>6</td>
<td>63</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>1029</td>
<td>104</td>
<td>5</td>
<td>64</td>
</tr>
</tbody>
</table>

*Single reference per disorder is shown. Full references are available in supplementary table 2 (available online only).

ADHD, attention deficit hyperactivity disorder; ASD, autism spectrum disorder.
HuGE set, this set included tested genes, not genes with significant findings, as associated genes (eg, 129 genes were associated with a single article,\textsuperscript{90} according to HuGE Navigator).

**DISCUSSION**

The main motivation for this study was that although there are many excellent sites\textsuperscript{21–93} providing detailed data on the human genome, it is cumbersome or not possible to retrieve the disease-associated genes together with supporting references from these sites. There are a few resources that provide disorder-targeted genes with references,\textsuperscript{54–94} but we found some issues related to references in such sites that cannot be addressed easily by the user as exemplified in the results section.

The innovation of our approach over existing tools is in the increased precision and that it is directly applicable within the context of statistical genetics and human genetic disorder research. Our method includes three formal steps—(1) extended query, (2) keyword filter, and (3) evaluation of abstract structure—to retrieve target disorder-specific, genetics-related articles. When compared to the mainstream data repositories such as HuGE Navigator, GAD, and OMIM, our three steps appear consistently to avoid the inclusion of non-genetics/non-research references and avoid mis-tagging genes/disorders.

Despite the encouraging results shown here, there are a number of limitations of our approach. First, as it focuses on extracting human genes and disorders, it will not accurately extract genes from model animal studies, for example zebrafish as an animal model for human fetal brain development.\textsuperscript{95, 96}

Second, because our tool uses a precision-based algorithm to extract genes, non-authoritative gene names/symbols that are not included in NCBI genes or HGNCG may not be properly matched to the correct gene symbols. Finally, our method searches titles and abstracts and therefore will not recover a gene association that is only mentioned in the main text. This could impact the sensitivity with GWAS that report many genes or loci in one article as a list in the main text.

We plan to expand our approach to the full text of the articles as future work. As expected from the BioCreative II task,\textsuperscript{31, 36} our rule-based algorithm successfully worked within the focused set of human genome research articles, and within concise data of titles and abstracts. However, statistical or hybrid entity recognition approaches may perform better in full text analysis, as shown in the BioCreative III task.\textsuperscript{36} We will examine this hypothesis with conditional random field-based models.\textsuperscript{41–42, 97, 98}

**CONCLUSION**

In this work, we introduced a novel PubMed extraction tool that can find and summarize research articles presenting evidence of gene–disorder associations. Comparison with existing resources demonstrated that our tool can cover more references in general and extract candidate genes with an accuracy comparable to manually curated sites. This application provides a fundamental basis for conducting cross-disorder analysis among related disorders, including solid evidence in the literature for every gene–disorder association. The overall candidate gene sets and supporting reference information are available at http://genehawk.hms.harvard.edu, and we plan to update result sets periodically as new publications come out.

**Contributors** DPW conceived the study. JYJ, DPW and TFD designed the study. JYJ, TFD and THN participated in implementation. All authors contributed to writing and approval of the final manuscript.

**Funding** This work was supported by the National Science Foundation grant nos 0543480 and 0640809 to DPW; and the National Institutes of Health grant no. LM009261 to DPW.

**Competing interests** None.

**Patient consent** Obtained.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** The overall candidate gene sets and supporting reference information are available online: http://genehawk.hms.harvard.edu.

**Open Access** This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/3.0/

**REFERENCES**


---

**Table 3** Categorization of ASD-specific articles

<table>
<thead>
<tr>
<th>Category</th>
<th># Ref.</th>
<th>Example ref.</th>
<th>Category</th>
<th># Ref.</th>
<th>Example ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviews</td>
<td>232</td>
<td>65</td>
<td>Common/GWAS</td>
<td>272</td>
<td>66</td>
</tr>
<tr>
<td>Case reports</td>
<td>135</td>
<td>67</td>
<td>Rare/CNV</td>
<td>139</td>
<td>68</td>
</tr>
<tr>
<td>Animal models</td>
<td>37</td>
<td>69</td>
<td>Exome sequencing</td>
<td>5</td>
<td>70</td>
</tr>
<tr>
<td>Gene expression</td>
<td>52</td>
<td>71</td>
<td>Blood/serum</td>
<td>23</td>
<td>72</td>
</tr>
</tbody>
</table>

*Single reference per category is shown. Full references are available in supplementary table 3 (available online only).

ASD, autism spectrum disorder; CNV, copy-number variation; GWAS, genome-wide association studies.

**Table 4** Top 10 candidate genes, ordered by collective article score for each gene

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Name</th>
<th>Score</th>
<th>#Ref.</th>
<th>Example references</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNTNAP2</td>
<td>Contactin-associated protein-like 2</td>
<td>56.23</td>
<td>34</td>
<td>24</td>
</tr>
<tr>
<td>FMR1</td>
<td>Fragile X mental retardation 1</td>
<td>50.10</td>
<td>68</td>
<td>75</td>
</tr>
<tr>
<td>SHANK3</td>
<td>SH3 and multiple ankyrin repeat domains 3</td>
<td>48.89</td>
<td>32</td>
<td>76</td>
</tr>
<tr>
<td>MET</td>
<td>Met proto-oncogene</td>
<td>47.18</td>
<td>19</td>
<td>77</td>
</tr>
<tr>
<td>SLC6A4</td>
<td>Neuronal transporter, serotonin</td>
<td>42.87</td>
<td>36</td>
<td>78</td>
</tr>
<tr>
<td>GABRB3</td>
<td>γ-Aminobutyric acid A receptor, subunit β 3</td>
<td>40.08</td>
<td>30</td>
<td>79</td>
</tr>
<tr>
<td>MECP2</td>
<td>Methyl CpG binding protein 2</td>
<td>39.64</td>
<td>47</td>
<td>80</td>
</tr>
<tr>
<td>PTEN</td>
<td>Phosphatase and tensin homolog</td>
<td>38.09</td>
<td>27</td>
<td>81</td>
</tr>
<tr>
<td>NRXN1</td>
<td>Neurexin 1</td>
<td>32.52</td>
<td>23</td>
<td>82</td>
</tr>
<tr>
<td>EN2</td>
<td>Engrailed 2</td>
<td>31.72</td>
<td>17</td>
<td>83</td>
</tr>
</tbody>
</table>

*Single reference per gene is shown. Full references are available in supplementary table 4 (available online only).


