BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 1 2010, pages 145–146
doi:10.1093/bioinformatics/btp618

Data and text mining

Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations

W. Yu*, M. Clyne, M. J. Khoury and M. Gwinn

Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, GA 30345, USA

Received on August 25, 2009; revised on October 7, 2009; accepted on October 25, 2009

Advance Access publication October 27, 2009

Associate Editor: Jonathan Wren

ABSTRACT

Summary: We developed web-based applications that encourage the exploration of the literature on human genetic associations by using a database that is continuously updated from PubMed. These applications provide user-friendly interfaces for searching summarized information on human genetic associations, using either genes or diseases as the starting point.

Availability: Phenopedia and Genopedia can be freely accessed at http://www.hugenavigator.net/HuGENavigator/startPagePhenoPedia.do and http://www.hugenavigator.net/HuGENavigator/startPagePedia.do, respectively.

Contact: wby0@cdc.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Advances in genomic technologies have dramatically boosted research on genetic associations (Kim and Misra, 2007), including genome-wide association studies (Neale and Purcell, 2008). Rapid growth in this field is reflected in the burgeoning number of related publications in public access databases such as PubMed (http://www.ncbi.nlm.nih.gov/pubmed/). Providing access to published information in an easy, comprehensive and systematic fashion is a critical first step in the synthesis and translation of genomic research data; however, information overload makes the retrieval, curation and presentation of such data an extremely challenging task. Since 2001, we have systematically collected and curated data on genetic association retrieved from PubMed, and deposited them in a database (Lin *et al*., 2006). We recently developed a screening program for genetic association literature that uses a machine learning technique called support vector machine for automatic classification. The new application significantly increased the recall, specificity and precision of screening (Yu *et al*., 2008a). Along with the new screening tool, the deployment of a new web-based system called HuGE Navigator for querying and filtering the data (Yu *et al*., 2008b) makes the database more robust, user-friendly and complete. Here, we present two extensions of HuGE Navigator, Phenopedia and Genopedia—integrated, web-based applications that display comprehensive summaries of published gene–disease associations, organized either by disease or by gene.

2 IMPLEMENTATION

As components of the integrated knowledge base on human genome epidemiology (HuGE Navigator) (Yu *et al*., 2008b), Phenopedia and Genopedia were built on J2EE technology (http://java.sun.com/javaee/) and on other Java open source frameworks such as Hibernate (http://www.hibernate.org/), Strut (http://struts.apache.org/), JChart (http://jcharts.krysalis.org/) and Google MAP API (http://www.google.com/apis/maps/documentation/). The database contains a curated collection of records retrieved weekly from PubMed since 2001. Each week, an automatic literature screening program (Yu *et al*., 2008a) screens PubMed for abstracts reporting gene–disease associations. A genetic epidemiologist selects abstracts meeting inclusion criteria and indexes them by gene, category and study type (Lin *et al*., 2006; Supplementary Table). Once staff of the National Center for Biotechnology Information (NCBI) has assigned Medical Subject Headings (MeSH) terms to abstracts in PubMed, they are retrieved using (NCBI) E-Utilities (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html) and used to index the records in the HuGE Navigator database. Disease terms in HuGE Navigator include all MeSH terms under the disease category in MeSH terminology (http://www.ncbi.nlm.nih.gov/sites/entrez?db=mesh). The metathesaurus in the Unified Medical Language System (Lindberg *et al*., 1993) is used as a lookup table for disease term synonyms. Entrez Gene records from the NCBI Entrez Gene database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gene) are used as standards for gene information. Data from the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa and Goto, 2000) are used to populate pathway information. The detailed schema for the HuGE Navigator database can be found in the paper by Yu *et al*. (2007).

3 FEATURES

Phenopedia provides a disease-centered view of genetic association studies. Information about genes studied in relation to a particular disease (e.g. stroke) or phenotype (e.g. hypertension) is summarized on the web page in tabular format. To perform a search, the user enters a disease term in the search box; the system maps the search text onto possible MeSH disease term(s), and the user selects from among all possible terms. Results of the search include: (i) the number of published genetic association studies, including the numbers of meta-analyses and genome-wide association studies (GWAS); (ii) the number of genes studied; (iii) the number of investigators (published authors) in the field; and (iv) temporal and

*To whom correspondence should be addressed.

geographic publication trends. A separate table displays a list of genes in descending order of the frequency with which they have been studied for association with the disease. For each gene, the table includes the total number of publications, as well as the numbers of meta-analyses, GWAS and gene–environment interaction studies; a link leads to a display of publication trends for the specific gene–disease association (Supplementary Fig. 1A and D). Each number in the table is a hyperlink that leads to a corresponding detailed information page. For example, the link for the number of publications links to results for the relevant query in HuGE Literature Finder, one of the other applications in HuGE Navigator.

We applied the method developed by Goh *et al.* (2007) to build a literature-based disease–gene network in which two diseases are ‘connected’ if they have been studied for association with the same gene(s). For a given disease, the display page lists each connected disease and the genes studied for association with both diseases (Supplementary Fig. 1B); data are shown in tabular form, without graphic representation. In addition, genes in the tables can be grouped according to pathways defined in KEGG (Supplementary Fig. 1C). The summary page provides links to some major disease-specific databases and published field synopses (Khoury *et al.*, 2009) if available.

Similarly, Genopedia provides a gene-centered summary view of genetic association studies. The system translates a gene name, gene symbol, gene alias or protein name entered by the user into a HUGO gene symbol. Genopedia displays information about diseases that have been studied in association with a given gene using a format similar to that described for Phenopedia (Supplementary Fig. 2A). A gene–disease network is also generated by defining two genes as ‘connected’ if they have been studied for association with the same disease(s) (Goh *et al.*, 2007). The list of connected genes is displayed along with the disease(s) that connect them (Supplementary Fig. 2B). Each search result page provides links to the foremost gene-centered databases, including NCBI Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gene), GeneCards (http://www.genecards.org/), PharmGKB (http://www.pharmgkb.org/) and ALFRED (http://alfred.med.yale.edu/). Genopedia has also become a major resource linkout for these major gene-centered databases.

As of June 2009, the HuGE Navigator database contains 2506 disease MeSH terms and 5456 genes. Phenopedia and Genopedia search results (a list of genes or a list of diseases) and whole datasets for the connections between genes and diseases are downloadable in a tab-delimited text file format. Both applications are components of HuGE Navigator, which allows navigation among all components as needed.

4 CONCLUSIONS

Our goal in developing Phenopedia and Genopedia was to provide researchers with quick and easy access to updated information on human genetic association studies, in order to facilitate knowledge synthesis, which is the first step in translating new knowledge gained from basic research to applications for clinical practice and public health (Khoury *et al.*, 2007). With these applications, we do not attempt to quantify specific genotype–phenotype associations; instead, we provide a starting point for systematic review and evaluation of associations by meta-analysis or other methods (Ioannidis *et al.*, 2008). Phenopedia and Genopedia serve as resources for the development of field synopses, which are regularly updated summaries of genetic associations in a particular field of research defined by a phenotype or family of genes (Khoury *et al.*, 2009). Our database differs in several key respects from the Online Mendelian Inheritance in Man (OMIM; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db¼OMIM) database, which focuses on rare disease-causing genetic variants and has only recently begun to include more common diseases. Our database systematically collects population-based genetic association studies dealing with common diseases, regardless of whether they report positive findings. Our applications display search results on the web page in tabular format, which is more efficient and user-friendly than the free plain text format used by OMIM. The automatic generation of hypothetical disease–gene networks and the integration of pathway information provide additional means for exploring hidden potential connections (Ekins *et al.*, 2007; Goh *et al.*, 2007). Currently, the association data in our database are indexed only at the gene level. We have experimented with extracting and displaying gene–disease association data (including published reference, phenotype, number of studies, number of cases, number of controls, contrast, effect size and heterogeneity) at the variant level for meta-analysis studies only (see example in Supplementary Fig. 1E). In our future work, we plan to collect gene variant-level information systematically and display it on the web accordingly, in table format. We also plan to create application programming interfaces or web services to facilitate integration with other systems.

Conflict of Interest: none declared

REFERENCES

Ekins,S. *et al.* (2007) Pathway mapping tools for analysis of high content data. *Methods Mol. Biol.*, **356**, 319–350.

Goh,K.I. *et al.* (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Ioannidis,J.P. *et al.* (2008) Assessment of cumulative evidence on genetic associations: interim guidelines. *Int. J. Epidemiol.*, **37**, 120–132.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Khoury,M.J. *et al.* (2007) The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? *Genet. Med.*, **9**, 665–674.

Khoury,M.J. *et al.* (2009) Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. *Am. J. Epidemiol.*, **170**, 269–279.

Kim,S. and Misra,A. (2007) SNP genotyping: technologies and biomedical applications. *Annu. Rev. Biomed. Eng.*, **9**, 289–320.

Lin,B.K. *et al.* (2006) Tracking the epidemiology of human genes in the literature: the HuGE Published Literature database. *Am. J. Epidemiol.*, **164**, 1–4.

Lindberg,D.A. *et al.* (1993) The Unified Medical Language System. *Methods Inf. Med.*, **32**, 281–291.

Neale,B.M. and Purcell,S. (2008) The positives, protocols, and perils of genome-wide association. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, **147B**, 1288–1294.

Yu,W. *et al.* (2007) An open source infrastructure for managing knowledge and finding potential collaborators in a domain-specific subset of PubMed, with an example from human genome epidemiology. *BMC Bioinformatics*, **8**, 436.

Yu,W. *et al.* (2008a) GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC Bioinformatics*, **9**, 205.

Yu,W. *et al.* (2008b) A navigator for human genome epidemiology. *Nat. Genet.*, **40**, 124–125.