American Journal of Epidemiology Advance Access published April 26, 2006

American Journal of Epidemiology

DOI: 10.1093/aje/kwj175

Practice of Epidemiology

Tracking the Epidemiology of Human Genes in the Literature: The HuGE Published Literature Database

Bruce K. Lin[1], Melinda Clyne[1], Matthew Walsh[2], Onnalee Gomez[3], Wei Yu[1], Marta Gwinn[1], and Muin J. Khoury[1]

[1] Office of Genomics and Disease Prevention, Coordinating Center for Health Promotion, Centers for Disease Control and Prevention, Atlanta, GA.
[2] Department of Population Health Sciences, University of Wisconsin, Madison, WI.
[3] Information Center, Coordinating Center for Health Information Systems, Centers for Disease Control and Prevention, Atlanta, GA.

Received for publication October 27, 2005; accepted for publication February 13, 2006.

Completion of the human genome sequence has inspired a new wave of epidemiologic studies on the prevalence of gene variants and their associations with diseases in human populations. In 2001, the Human Genome Epidemiology (HuGE) Network launched the HuGE Published Literature database (HuGE Pub Lit), a searchable, online knowledge base of published, population-based epidemiologic studies of human genes. The database contains links to PubMed articles and can be searched by gene, disease, interacting factor, type of study design or analysis, or any combination of terms in these categories. The search output contains a link to each identified article, along with a table summarizing key features of the reported study. As of September 6, 2005, some 17,665 articles were indexed in the database. Most described gene-disease associations (86%); fewer evaluated gene-gene or gene-environment interactions (17%), the prevalence of gene variants (10%), or genetic tests (3%). Although not comprehensive, this database is a unique tool for epidemiologic researchers and others concerned with the role of genetic variation in population health. Here, the authors provide an overview of the database and its characteristics and uses.

databases, genetic; epidemiology, molecular; genetics; genetics, population; genome, human; genomics; PubMed

Abbreviations: HuGE, Human Genome Epidemiology; HuGE Net, HuGE Network; HuGE Pub Lit, HuGE Published Literature database.

Advances in genomics have encouraged the use of epidemiologic methods to measure the prevalence of genetic variants in populations, to characterize gene-disease associations, to identify gene-gene and gene-environment interactions, and to evaluate genetic tests. The epidemiologic approach to the human genome—or Human Genome Epidemiology (HuGE)—covers the spectrum from gene discovery to clinical and public health practice (1). The HuGE Network (HuGE Net), established in 1998, is an open, global collaboration of individuals and organizations from clinical medicine, public health, and other fields who are committed to developing and disseminating epidemiologic information on the human genome (2, 3). HuGE Net promotes the systematic review and synthesis of this information in HuGE

Correspondence to Dr. Marta Gwinn, Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, Mailstop K89, 4770 Buford Highway, Atlanta, GA 30341 (e-mail: mgwinn@cdc.gov).

reviews, which are published in partnership with several journals, including the *American Journal of Epidemiology* (3).

In 2001, HuGE Net launched the HuGE Published Literature database (HuGE Pub Lit), a continually updated and accessible knowledge base on the World Wide Web that tracks the growing published literature of human genome epidemiologic studies. HuGE Pub Lit offers a starting point for assembling articles for meta-analysis, highlighting research gaps, suggesting applied research questions, and identifying potential collaborators.

In this paper, we describe the contents of HuGE Pub Lit, how eligible articles are identified, and the database's current utility and limitations in tracking the results of human genome epidemiologic research.

OVERVIEW OF THE DATABASE

HuGE Pub Lit consists of links to PubMed entries (4), extracted weekly since October 21, 2000. As of September 6, 2005, 17,665 articles were indexed in the database, referencing 2,008 genes (using the standard nomenclature established by the Human Genome Organization) (5), 2,074 diseases and health outcomes, and 660 interacting personal or environmental factors; for an update, refer to ''Note added in proof.'' Articles are included if they present data on any one of five topics: genotype prevalence, gene-disease associations, gene-gene interactions, gene-environment interactions, or the validity or utility of genetic tests (2). Gene discovery articles (e.g., those describing linkage analysis and gene mapping in high-risk families or selected groups) are excluded; however, articles are examined on a case-by-case basis, because it is becoming more common for gene discovery articles to report additional, corroborating association studies (6) and for population-based studies to include related persons (7). Inclusion and exclusion criteria have evolved since 2001, reflecting advances in human genome epidemiologic research methods (table 1).

Detailed instructions for querying the database are provided on the website (8). Query results display the number of relevant articles from HuGE Pub Lit and a list of citations. Each citation provides links to the original PubMed abstract through the National Library of Medicine's website and a brief annotation of the relevant gene, disease or outcome, personal or environmental factor, and topic.

METHODS FOR EXTRACTING HuGE ARTICLES FROM PubMed

Centers for Disease Control and Prevention research staff extract relevant articles from PubMed into HuGE Pub Lit each week using a combination of computerized and manual processes. First, all titles, abstracts, and PubMed fields are searched by use of a complex query based on a combination of Medical Subject Headings' (MeSH; National Library of Medicine, Bethesda, Maryland) terms and selected text words. The search query selects on average 1,800 articles from over 20,000 new entries in PubMed each week. The curator (M. Clyne) reviews each title and abstract to decide whether the study will be included in the database and to assign index terms. Citations for the approximately 100 articles that meet HuGE criteria are posted every Thursday as part of the Centers for Disease Control and Prevention online *Genomics and Health Weekly Update* (9). Quality checks are performed within the finished database.

TABLE 1. Criteria used to select articles from PubMed for HuGE Pub Lit*

Inclusion criteria
Published since October 1, 2000
English-language abstract (full text may be in any language)
Human study population
Genotype measured or inferred at one or more loci
Epidemiologic study design: cohort, case-control, case only, clinical trial
Population-based analysis: genotype prevalence, association with disease, interactions with other genes or environmental factors, validity, or utility of genetic test
Exclusion criteria
No human data (e.g., study of gene function in an animal model)
No genotype (e.g., association with chromosome region only)
No population-based analysis (e.g., linkage study in high-risk family, case report)

* HuGE Pub Lit, Human Genome Epidemiology Published Literature database.

The process for identifying HuGE published literature is currently restricted to the contents of PubMed. The sensitivity of the process is limited by the information content of titles, abstracts, and other PubMed fields for individual articles. Unless the full text of an article is available, the curator may omit it if the abstract does not include relevant terms. Uploading to HuGE Pub Lit also lags for articles that require further review because the abstract is absent, incomplete, or in a foreign language. As the HuGE literature grows, so does the number of potentially eligible articles that must be reviewed.

To assess the sensitivity of the routine extraction process, an independent abstractor reviewed all 4,556 articles that had been entered into PubMed on six randomly selected dates between October 2000 and December 2002. After review by two other researchers, the narrowed list included 39 eligible articles; of these, 31 had been previously identified by the routine extraction process, for a sensitivity of 80 percent. Although the search query used during that period had identified all 39 articles, eight had been incorrectly excluded by the curator. None of the articles previously included in HuGE Pub Lit were excluded on review, for a specificity of 100 percent.

DESCRIBING THE PUBLISHED LITERATURE ON HUMAN GENOME EPIDEMIOLOGY

The contents of HuGE Pub Lit reflect a steady increase in the number of relevant publications, from 3,024 in 2001 to 5,247 in 2004. Most articles in HuGE Pub Lit present data on

TABLE 2. Ten most commonly referenced health conditions in HuGE Pub Lit*

Health condition	No. of references
Breast cancer	614
Alzheimer's disease	510
Diabetes, type 2	510
Atherosclerosis, coronary	480
Schizophrenia	470
Hypertension	412
Colorectal cancer	311
Myocardial infarction	301
Obesity	299
Lung cancer	298

* HuGE Pub Lit, Human Genome Epidemiology Published Literature database (as of September 6, 2005).

gene-disease associations (86 percent); 17 percent describe gene-gene or gene-environment interactions, and only 10 percent describe genotype prevalence. HuGE Pub Lit lends itself to additional descriptive analyses of the published literature in human genome epidemiology. For example, the 10 most commonly referenced genes are involved in pathways that affect multiple diseases: four (*HLA-DRB1, HLA-DQB1, TNF, IL10*) are involved in the body's immune or inflammatory responses, two (*GSTM1, GSTT1*) are involved in the detoxification of xenobiotics, two (*APOE, MTHFR*) are important in nutrient absorption or metabolism, and two (*ACE, F5*) are involved in circulation and hemostasis. All of the 10 most commonly referenced health conditions are chronic conditions of adults (table 2). Smoking (tobacco) is by far the most commonly studied interacting factor, referenced in nearly 500 articles.

As an example, we describe the contents of HuGE Pub Lit for the relation between variants in *MTHFR* and congenital anomalies (table 3). Of the 17,665 publications in the database as of September 6, 2005, 727 (4 percent) dealt with *MTHFR*. Among the 50 publications describing *MTHFR* variants in relation to congenital anomalies, 11 also addressed gene-gene or gene-environment interactions. So far, 20 systematic reviews (including three HuGE reviews) have been published on epidemiologic associations with *MTHFR;* one of these examined the association with congenital anomalies.

TABLE 3. Example of search in HuGE Pub Lit*

Type of publication	All genes	*MTHFR*	*MTHFR*/ congenital anomalies
All publications	17,665	729	50
Publications describing gene-gene or gene-environment interaction	2,966	173	11
Systematic reviews	308	20	1
HuGE* reviews	39	3	1

* HuGE Pub Lit, Human Genome Epidemiology Published Literature database (as of September 6, 2005); HuGE, Human Genome Epidemiology.

DISCUSSION

HuGE Pub Lit fills a specific niche as an online information resource for epidemiologic and translation research. Currently, a vast and complex array of genetic databases and medical genetics resources and sites devoted to genomics can be accessed on the Internet (10–15). Most of these databases provide either biologic data for genetics researchers or information about genetic disorders for clinicians. Of the many Web resources that have cataloged genes, sequences, or variants, a small subset also offers information on phenotypic associations. For example, Online Mendelian Inheritance in Man (OMIM), one of the first databases to catalog genes, includes descriptions of associated diseases and outcomes (14, 15). The National Institute on Aging's Genetic Association Database (16) contains gene-based records for association studies along with links to genetic mapping and sequence databases. Among the unique features of HuGE Pub Lit are its focus on population-based epidemiologic research; indexing by health outcomes, interacting factors, and epidemiologic terms; weekly updating of content; and curation by an epidemiologist.

Although the HuGE Net website is widely used (more than 160,000 visits during the latest 12-month period), the maintenance of HuGE Pub Lit presents numerous challenges. Currently, the most urgent need is for a more efficient and reproducible method of identifying articles of interest. The current process is effective but subjective and time consuming, occupying one epidemiologist curator full time. We are currently collaborating with computer scientists at the Georgia Institute of Technology to explore the development of text-mining tools, which have been used in related fields (e.g., pharmacogenomics) (17), and our experience thus far shows some promising results (18). Of course, review of database content by subject matter experts will remain an essential element of quality control.

A standard vocabulary for outcomes, interacting factors, and epidemiologic concepts is also needed to enhance the interoperability among databases and exchangeability among users. Although such standards have been developed for genes (5), they are still evolving for other types of data. Ultimately, the use of a standard vocabulary, such as that found within the Unified Medical Language System (UMLS) (19), will make it possible to index articles in a more systematic, standard, and automatic fashion and allow more natural searching (e.g., by using free text). Standard vocabulary will also facilitate searching that navigates smoothly from the epidemiologic literature to relevant research in other disciplines, such as molecular biology and toxicology.

Currently, HuGE Pub Lit does not capture studies published in journals that are not included in the PubMed database. Some studies could be identified from other publication databases (e.g., EMBASE) (20), but others are unpublished or presented only at scientific meetings (so-called ''gray literature'') (21). Capturing information from these sources is particularly important to reduce the influence of

publication bias in meta-analysis. We have recently proposed a more comprehensive approach to the synthesis of evidence on gene-disease association that calls for coordinated efforts by investigators and additional avenues for publishing the results of well-conducted studies, regardless of the findings (22).

We encourage researchers to consider using HuGE Pub Lit as a quick search tool for identifying published epidemiologic data and as an adjunct to other sources when developing systematic reviews (23–25). We invite curators of similar genetic association databases to collaborate with us in making our systems interoperable and in developing the informatics tools to support research synthesis in human genome epidemiology (26). We hope that HuGE Pub Lit will provide a useful starting point for building the knowledge base on human genetic variation and health, which is the first step in translating human gene discoveries into health care and disease prevention.

Note added in proof: *Since the acceptance of this paper, the number of articles in the HuGE Pub Lit database has increased substantially. As of April 21, 2006, the database included 20,711 articles.*

ACKNOWLEDGMENTS

The authors thank Paula Yoon, Ajay Yesupriya, Lori Durand, and Alex Charles for their contributions to HuGE Pub Lit.

Conflict of interest: none declared.

REFERENCES

1. Khoury MJ, Little J, Burke W, eds. Human genome epidemiology: a scientific foundation for using genetic information to improve health and prevent disease. New York, NY: Oxford University Press, 2004.
2. Khoury MJ, Dorman JS. The Human Genome Epidemiology Network. Am J Epidemiol 1998;148:1–3.
3. Human Genome Epidemiology Network. Atlanta, GA: Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, 2005. (http://www.cdc.gov/genomics/hugenet/default.htm). Accessed August 20, 2005.
4. PubMed. Bethesda, MD: National Library of Medicine, 2005. (http://www.ncbi.nlm.nih.gov/entrez). Accessed August 31, 2005.
5. HUGO Gene Nomenclature Committee. London, United Kingdom: The Human Genome Organisation, 2004. (http://www.gene.ucl.ac.uk/nomenclature). Accessed August 31, 2005.
6. Ogura Y, Bonen DK, Inohara N, et al. A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. Nature 2001;411:603–6.
7. Karamohamed S, Demissie S, Volcjak J, et al. Polymorphisms in the insulin-degrading enzyme gene are associated with type 2 diabetes in men from the NHLBI Framingham Heart Study. Diabetes 2003;52:1562–7.
8. About the HuGE Published Literature database. Atlanta, GA: Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, 2005. (http://www.cdc.gov/genomics/search/aboutHPLD.htm). Accessed September 6, 2005.
9. Genomics and health weekly update. Vol 15, no. 7. Atlanta, GA: Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, 2005. (http://www.cdc.gov/genomics/update/current.htm). Accessed August 20, 2005.
10. Guttmacher AE. Human genetics on the Web. Annu Rev Genomics Hum Genet 2001;2:213–33.
11. Marks AD, Yoon PW. Human Genome Epidemiology information (HuGE) on the Internet: current resources and future prospects. Atlanta, GA: Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, 2005. (http://www.cdc.gov/genomics/hugenet/genData.htm). Accessed August 31, 2005.
12. Thomas S, curator. MetaDB: a metadatabase for the biological sciences. Urbana-Champaign, IL: University of Illinois, 2005. (http://www.neurotransmitter.net/metadb/metadb.php). Accessed August 31, 2005.
13. Human Genome Variation Society. Fitzroy, Victoria, Australia: Genomic Disorders Research Centre, St.Vincent's Hospital Melbourne, 2005. (http://www.hgvs.org). Accessed September 6, 2005.
14. OMIM—Online Mendelian Inheritance in Man. Bethesda, MD: National Center for Biotechnology Information, 2005. (http://www3.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM). Accessed August 31, 2005.
15. Hamosh A, Scott AF, Amberger JS, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005;33:D514–17.
16. Genetic Association Database. Bethesda, MD: National Institute on Aging, National Institutes of Health, 2005. (http://geneticassociationdb.nih.gov/). Accessed August 31, 2005.
17. Rubin DL, Thorn CF, Klein TE, et al. A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge. J Am Med Inform Assoc 2005;12:121–9.
18. Polavarapu N, Navathe SB, Ramnarayanan R, et al. Investigation into biomedical literature classification using support vector machines. (Abstract). Presented at the 2005 IEEE Computer Society Bioinformatics Conference, Stanford, California, August 8–11, 2005.
19. Unified Medical Language System. Bethesda, MD: National Library of Medicine, 2005. (http://www.nlm.nih.gov/research/umls/). Accessed August 20, 2005.
20. EMBASE Excerpta Medica. New York, NY: Elsevier, 2005. (http://www.elsevier.com/wps/find/bibliographicdatabasedescription.cws_home/523328/description). Accessed August 31, 2005.
21. Alpi KM. Expert searching in public health. J Med Libr Assoc 2005;93:97–103.
22. Ioannidis JPA, Gwinn M, Little J, et al. A road map for efficient and reliable human genome epidemiology. Nat Genet 2006;38:3–5.
23. Lai R, Crevier L, Thabane L. Genetic polymorphisms of glutathione *S*-transferases and the risk of adult brain tumors: a meta-analysis. Cancer Epidemiol Biomarkers Prev 2005;14:1784–90.
24. Khoury MJ, Little J. Human genome epidemiologic reviews: the beginning of something HuGE. Am J Epidemiol 2000;141:839–45.
25. Little J, Khoury MJ, Bradley L, et al. The Human Genome Project is complete. How do we develop a handle for the pump? Am J Epidemiol 2003;157:667–73.
26. Bracken MB. Genomic epidemiology of complex diseases: the need for an electronic evidence-based approach to research synthesis. Am J Epidemiol 2005;162:297–301.