A navigator for human genome epidemiology

To the Editor:

Recent successes in large-scale genetic association studies call for renewed attention to integrating research results, not only among studies, but across disciplines[1]. At the molecular level, genetic polymorphisms provide a starting point for investigating the functions of complex biological systems. At the population level, epidemiologists can begin to use data on genetic variation, associations and interactions to interpret population attributable fractions and estimate the potential health impact of genetically directed interventions[2]. Publicly available genetic sequence databases have demonstrated their value in accelerating the Human Genome Project and advancing the field of molecular genetics; newer efforts, such as dbGaP and CGEMS, are now beginning to make genotype-phenotype data broadly available to the scientific community[3].

The published scientific literature also reflects rapid growth in studies of human genetic factors in relation to health and disease. Since 2001, the Human Genome Epidemiology Network (HuGENet) has maintained a database of published, population-based epidemiologic studies of human genes extracted from PubMed[4]. We recently replaced our PubMed search strategy with a new approach using machine learning, which has reduced manual effort and increased both the sensitivity and specificity of screening. Our curator updates the database weekly with articles newly added to PubMed and assigns to them one or more study types (for example, observational study, meta-analysis or genome-wide association study) and data categories (for example, gene-disease association, gene-environment interaction or pharmacogenomics). Each article is indexed in the database with MeSH terms (using the MeSH hierarchical structure) and gene information from the National Center for Bioinformatics (NCBI) Entrez Gene database. As of November 2007, the database has indexed more than 30,000 articles, referencing more than 3,000 genes and nearly 2,000 disease terms (**Table 1**). Most articles (80%) describe genetic associations. Approximately 20% of all articles were published in 2007, including 68 of 82 genome-wide association studies.

To make this database more accessible and useful to interdisciplinary researchers, we have developed an integrated set of applications known collectively as the HuGE Navigator (http://www.hugenavigator.net). Using PubMed abstracts as the core data source, we have developed data and text mining algorithms to create a knowledge base for exploring genetic associations, candidate gene selection and investigator networks. Genetic information can be displayed whenever needed from major gene-centered databases (for example, Entrez Gene, SwissProt, OMIM and GeneCards), as well as from databases of genetic variation and prevalence (for example, dbSNP and HapMap Project), pathways (for example, CGAP, KEGG and BioCarta), and other aspects (for example, Gene Ontology and Gene Clinics). The HuGE Navigator is constructed according to the principles of open source, standardization, interoperability and extensibility, so that new applications can be easily incorporated[5]. Currently, the HuGE Navigator allows users to navigate and search the database in an integrated manner by using the six applications discussed below.

The HuGE Literature Finder is a search engine for finding published literature on human genome epidemiology, including genetic association studies. The search query can include disease terms, environmental factors, genes, or author names and affiliations. The search results can be further refined by using filtering features, including disease, gene, category, study type, author, year, journal, and country. The filtering process can be performed indefinitely until the desired result is obtained. The results (PubMed IDs) can be exported to the PubMed Web site for further exploration and downloading to bibliographic software.

The HuGE Investigator Browser is a search engine for finding investigators or collaborators on the basis of research interests, such as diseases, risk factors, or genes. We extract investigator data by using an accessory utility that automatically parses the affiliation data provided by PubMed[6].

GeneSelectAssist is a search tool for finding possible candidate genes associated with the subject of interest. Search terms can include diseases and exposures. GeneSelectAssist selects and prioritizes genes on the basis of genetic association studies in the HuGE Navigator database, as well as other PubMed abstracts, and evidence from animal models in the NCBI Entrez Gene database.

HuGE Watch is a tool for tracking the evolution of human genome epidemiology research dynamically, on the basis of the literature database. It allows users to view temporal trends in publication by gene, disease, and number of investigators, as well as by the geographic distribution of authors.

HuGEpedia is an online encyclopedia that summarizes research on gene-disease associations. We are currently developing a system for extracting data from meta-analyses and published genome-wide associations that will form the basis for a disease-specific synopsis written by domain experts. HuGEpedia can be searched by gene or disease.

HuGE Risk Translator is a tool that assesses the validity of genetic variants for predicting health outcomes by calculating epidemiologic measures such as population attributable risk, sensitivity, specificity and positive and negative predictive values.

The HuGE Navigator offers a new way to navigate and mine the growing scientific literature on human gene-disease associations and related data in human genome epidemiology. As an interconnected system of applications that users can enter by using genes, diseases, or risk factors as the starting point, HuGE Navigator provides a potential bridge between epidemiologic and genetic research domains. Disease and gene names are mapped to standardized vocabularies, so investigators can use their preferred terms to query the knowledge base. By linking to disease-specific databases, such as AlzGene[7], HuGE Navigator aims to be the vehicle for navigating the 'network of networks' of investigators now working to

Table 1 HuGE Navigator database contents

Study type (category)[a]	Number of publications
Gene-disease association	26,638
Pharmacogenomics	1,802
Meta-analysis	593
HuGE review	61
Genome-wide association	80
All articles	30,490

[a]Articles can be assigned more than one study type.

untangle the complex relationships among genetic and environmental factors that underlie human disease[8].

Wei Yu, Marta Gwinn, Melinda Clyne, Ajay Yesupriya & Muin J Khoury

National Office of Public Health Genomics Centers for Disease Control and Prevention, Atlanta, Georgia 30309, USA. Correspondence should be addressed to W.Y. (wby0@cdc.gov).

1. Frayling, T.M. & McCarthy, M.I. *Diabetologia* **50**, 2229–2233 (2007).
2. Khoury, M.J. *et al.* 1–524 (Oxford University Press, New York, New York, 2004).
3. Mailman, M.D. *et al. Nat. Genet.* **39**, 1181–1186 (2007).
4. Lin, B.K. *et al. Am. J. Epidemiol.* **164**, 1–4 (2006).
5. Yu, W. *et al. BMC Med. Inform. Decis. Mak.* **20**, 17 (2007).
6. Yu, W. *et al. BMC Bioinformatics* **9**, 436 (2007).
7. Bertram, L., McQueen, M.B., Mullin, K., Blacker, D. & Tanzi, R.E. *Nat. Genet.* **39**, 17–23 (2007).
8. Ioannidis, J.P. *et al. Nat. Genet.* **38**, 3–5 (2006).