Along with the new screening tool, the deployment of a new web-based system called HuGE Navigator for querying and filtering the retrieval, curation and presentation of such data an extremely critical first step in the synthesis and translation of related publications in public access databases such as 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/query.fcgi?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/static/entrezgene_help.html) and used to deposit 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.
HuGE Navigator, which allows navigation among all components in a tab-delimited text file format. Both applications are components of the integration of pathway information provide additional means for exploring hidden potential connections (Ekins 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., 388, 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 Miesa,A. (2007) SNP genotyping: technologies and biomedical applications. Annu. Rev. Biomed. Eng., 9, 289–320.

Lee,D.B. 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 studies. Am. J. Med. Genet. B Neuropsychiatr. Genet., 147B, 1286–1294.

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) GApCreeper: 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.