European Journal of Human Genetics (2011) 19, 928–930

www.nature.com/ejhg

SHORT REPORT

Cancer GAMAdb: database of cancer genetic associations from meta-analyses and genome-wide association studies

Sheri D Schully*[,1,5], Wei Yu[2,5], Victoria McCallum[3], Camilla B Benedicto[1], Linda M Dong[4], Anja Wulf[2], Melinda Clyne[2] and Muin J Khoury[1,2]

In the field of cancer, genetic association studies are among the most active and well-funded research areas, and have produced hundreds of genetic associations, especially in the genome-wide association studies (GWAS) era. Knowledge synthesis of these discoveries is the first critical step in translating the rapidly emerging data from cancer genetic association research into potential applications for clinical practice. To facilitate the effort of translational research on cancer genetics, we have developed a continually updated database named Cancer Genome-wide Association and Meta Analyses database that contains key descriptive characteristics of each genetic association extracted from published GWAS and meta-analyses relevant to cancer risk. Here we describe the design and development of this tool with the aim of aiding the cancer research community to quickly obtain the current updated status in cancer genetic association studies.
European Journal of Human Genetics (2011) **19,** 928–930; doi:10.1038/ejhg.2011.53; published online 13 April 2011

Keywords: cancer; meta-analyses; pooled analyses; GWAS

INTRODUCTION

With advances in high throughput genotyping technologies,[1] the number of genetic association studies has increased at an unprecedented pace over the past decade. The systematic review of such studies, especially meta-analyses across multiple studies, has been recommended to minimize false-positive associations and as a tool to assess the credibility of the findings.[2] Genome-wide association studies (GWAS) have recently emerged as a powerful tool to find many novel genetic associations that the traditional candidate gene approach has failed to discover.[3] In the field of cancer, genetic association studies are among the most active and well-funded research areas and have produced hundreds of genetic associations, especially in the GWAS era. Although we do realize the limitations of each of the individual studies (ie, publication bias), meta-analysis, including heterogeneity testing, still can provide valuable information to genetic epidemiology researchers to supplement GWAS. Knowledge synthesis of these discoveries is the first critical step in translating the rapidly emerging data from cancer genetic association research into potential applications for clinical practice,[4] and is also important for basic scientists as they can build on these discoveries. To facilitate the effort of translational research on cancer genetics, we have developed a continually updated database named Cancer Genome-wide Association and Meta Analyses database (Cancer GAMAdb) that contains key descriptive characteristics of each genetic association extracted from published GWAS and meta-analyses relevant to cancer risk. Here we describe the design and development of this tool (http://www.hugenavigator.net/CancerGEMKB/caIntegratorStartPage.do). Our aim is to help the cancer research community to quickly obtain the current updated status in cancer genetic association studies and readily retrieve relevant information in a highly integrated manner. The database is supported as a joint venture between the National Cancer Institute's Division of Cancer Control and Population Sciences and the Centers for Disease Control and Prevention's Office of Public Health Genomics.

IMPLEMENTATION

The Cancer GAMAdb catalogs published GWAS and meta- and pooled analyses that have evaluated the association between genetic polymorphisms and cancer risk since 1 January 2000. The methodology used in creating this robust database can be seen in Figure 1. To efficiently retrieve the published genetic association articles from PubMed, a computerized text mining search algorithm with high sensitivity (97.5%) and specificity (98.3%),[5] combined with follow-up manual curation, is used to find genetic association articles from PubMed as part of a published literature database screening process in the Human Genome Epidemiology (HuGE) Navigator.[6] Among the HuGE literature repository, articles are eligible for inclusion if they meet the following criteria: (1) evaluate cancer risk as the outcome, (2) represent a GWAS study, meta-, or pooled analyses with aggregated estimates of effect, and (3) are published in English. The curator flags PubMed abstracts by 'meta-analysis', 'pooled analysis', or 'genome-wide association' if the articles fall within the inclusion

[1]Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, USA; [2]Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, GA, USA; [3]Office of Workforce Development, National Cancer Institute, Bethesda, MD, USA; [4]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA
*Correspondence: Dr SD Schully, Division of Cancer Control and Population Sciences, National Cancer Institute, 6130 Executive Boulevard, Suite 5124 MSC 7393, Bethesda, MD 20892, USA. Tel: +1 301 435 4911; Fax: +1 301 435 5477; E-mail: schullys@mail.nih.gov
[5]These authors contributed equally to this work.
Received 6 September 2010; revised 2 February 2011; accepted 11 February 2011; published online 13 April 2011

criteria. As a starting point, we used a previously published dataset by Dong *et al*,[7] which included meta-analyses and pooled analyses found in PubMed that evaluated the relationship between genetic polymorphisms and cancer risk through 15 March 2008. We also review relevant articles in the online NIH GWAS Catalog (http://www.genome.gov/26525384) as a quality check in case any GWAS articles have been overlooked. Data elements extracted from each full text article include cancer site, the gene and variant names, risk phenotype or allele, risk estimates (odds ratios or relative risk), 95% confidence intervals, ethnicity or gender (when applicable), minor allelic frequency (when applicable), number of studies, number of cases and controls, *P*-values, tests for heterogeneity, tests of publication bias, type of platform used (if GWAS), gene–environment interactions (if applicable), study replication (if GWAS), copy number variation (if applicable), study type (candidate, GWAS, or clinical trial), and analysis type (meta, pooled, or consortia). Random-effect estimates

Figure 1 Workflow of the methodology use to create the Cancer GAMAdb.

Cancer GAMAdb

- **Cancer** Genome-wide **A**ssociation and **M**eta **A**nalyses **D**atabase -

[Database Statistics] Home | About | Search Instructions | FAQs

Search Summary for bladder cancer Go Clear

Query Trace: bladder cancer[orginal query]>>Bladder Cancer[Phenotype]

HuGE Literature | UCSC Gene Track | UCSC Variant Track | Download

Search Results (Found a total of **222 records**) record 1 - 25 >> Sorted by: Phenotype Order: Ascending

- To fine-tune the query results, use these filter functions -

Phenotype | Variant | Gene | Publication | Author | Journal | Year | Study Type | G-E Interaction

Phenotype	Reference	Variant	Gene/Region	Contrast	#Studies	#Case	#Control	OR [95% CI]	p-value	Type*	Detail
Bladder Cancer	Zhang,2010 Mol Biol Rep	GSTM1 null	GSTM1	null	2	NA	NA	1.59 [0.99-2.57] (Asian, Fixed-effects model)	0.06	C	Detail
Bladder Cancer	Xu,2010 Mol Biol Rep	rs1800795 [common_name]	IL-6	CC v GC	2	580	643	4.33 [1.93-9.71]	<0.001	C	Detail
Bladder Cancer	Xu,2010 Mol Biol Rep	rs1800795 [common_name]	IL-6	CC v GG	2	580	643	2.81 [1.39-5.68]	0.004	C	Detail
Bladder Cancer	Xu,2010 Mol Biol Rep	rs1800795 [common_name]	IL-6	GC v GG	2	580	643	0.65 [0.37-1.12]	0.122	C	Detail
Bladder Cancer	Xu,2010 Mol Biol Rep	rs1800795 [common_name]	IL-6	GC + CC v GG	2	580	643	1.08 [0.69-1.70]	0.74	C	Detail
Bladder Cancer	Xu,2010 Mol Biol Rep	rs1800795 [common_name]	IL-6	CC v GG + GC	2	580	643	2.19 [1.32-3.64]	0.003	C	Detail
Bladder Cancer	Zhang,2010 Mol Biol Rep	GSTM1 null	GSTM1	null	4	507	818	1.6 [1.27-2.01] (Asian)	NR	C	Detail

Figure 2 Screenshot of the search for 'bladder cancer'.

from meta-analyses were used, unless the paper included only fixed-effect estimates. Significant associations from GWAS are recorded based on the NIH GWAS Catalog criteria (http://www.genome.gov/27529028). For the standardization of the cancer phenotypes, gene names, and variant names, we manually code phenotypes with a Unified Medical Language System (UMLS) unique identifier, gene names with the Human Genome Organisation gene symbol and National Center for Biotechnology Information (NCBI) Entrez Gene GeneID, and RefSNP accession ID (rs numbers) for the variant names if they are available. On the use of the UMLS Metathesaurus (http://www.nlm.nih.gov/research/umls/), NCBI Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gene), and Variant Name Mapper (http://www.hugenavigator.net/HuGENavigator/startPageMapper.do) as reference sources, the database offers a robust search capacity with a user-friendly web interface in a free-text search manner (Figure 2).

FEATURES

Cancer GAMAdb also provides analytic functionalities on the dataset selected. Filter features by nine key elements (ie, phenotype, gene, variant, publication, author, journal, year, study type, gene–environment interaction) on the retrieved records allow users to perform quick descriptive analyses on the associations of interest while undergoing the dataset search. The University of California Santa Cruz (UCSC) Genome Browser custom tracks are dynamically generated at gene and variant levels based on user's selected dataset, and a user may subsequently use all functionalities and information in the default tracks from the UCSC Genome Browser (http://genome.ucsc.edu/). As a component of HuGE Navigator, Cancer GAMAdb contains many key dynamic links to other components in the knowledge base where many more disease-specific or gene-central information can be retrieved, such as dynamic linkage from phenotypes to phenopedia and from genes to genopedia.[8] The primary research articles may be easily obtained by clicking the HuGE Literature button for the given search term. In addition, all data including the datasets in any selection steps are downloadable in a text format. The database statistics page dynamically outlines a comprehensive descriptive view on cancer genetic association research, including count numbers and graphs of temporal trends for phenotype/variant associations, publications, genes studied, variants studied, and phenotypes reported. In addition, a series of top 10 lists are generated and displayed in web tables with regards to variant, gene, phenotype, author, or journal, itemized by GWAS and meta-analyses. As of 28 January 2011, the database contains 5354 reported cumulative genetic associations relevant to cancer risk from 599 publications, including 504 meta-analyses and 95 GWAS (see Table 1). The summary data in the statistic page indicates that, in terms of cancer genetic risk, breast cancer is the most studied disease in both meta-analyses and GWAS; *GSTM1* null allele is the most commonly studied variant among meta-analyses and may be associated with 15 different cancer phenotypes; and rs6983267 (a SNP on 8q24) is the top statistically significant variant that has appeared in eight GWAS publications.

Table 1 Cancer GAMAdb content (as of 28 January 2011)

Category	*Count*
Association	6180 (total), 5647 (meta-analysis), 519 (GWAS)
Publication	599 (total), 504 (meta-analysis), 95 (GWAS)
Gene	489 (total), 254 (meta-analysis), 268 (GWAS)
Variant	1000 (total), 718 (meta-analysis), 337 (GWAS)
Phenotype	67 (total), 58 (meta-analysis), 33 (GWAS)

Abbreviations: GAMAdb, Genome-wide Association and Meta Analyses database; GWAS, Genome-wide association study.

CONCLUSION

Cancer GAMAdb is continually updated to accurately track the rapid progress in cancer genetic association research, and offers a valuable bioinformatics tool for cancer researchers and clinical practitioners to quickly obtain current information on the latest association studies and the most recent status of the research. To our best knowledge, Cancer GAMAdb is the first online searchable database that deposits cancer genetic association information from published meta- and pooled analyses and GWAS studies. There is no other resource that combines data generated by these methodologies. The Cancer GAMAdb is a key step to knowledge synthesis of key cancer genetic epidemiology findings. This database also is an indispensable component of the integrated toolset in the HuGE Navigator knowledge base for cancer knowledge. It enhances phenopedia,[8] which displays a comprehensive summary web table listing all possible association genes and the numbers of published articles for each gene, by allowing users to quickly retrieve variant-level associations from meta-analyses or GWAS by linking to the Cancer GAMAdb. The same navigation can start from Cancer GAMAdb leading to phenopedia. So far, there are only a few disease-specific genetic association databases available, such as AlzGene (http://www.alzgene.org/), which have collected and extracted data from all primary published literature in the field. Although such information is extremely valuable, extraction of detailed information from the published literature is always labor intensive and time consuming. Cancer GAMAdb creates a new way to capture genetic associations by cataloging only summarized genetic associations from meta-analyses, pooled analyses, and significant findings from GWAS studies. This valuable tool could significantly reduce efforts to create and maintain such disease-specific databases while making important information in the field easily accessible and available to the research community.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DISCLAIMER

The findings and conclusions in this report are those of the authors and do not necessarily reflect the views of the Department of Health and Human Services.

1 Panoutsopoulou K, Zeggini E: Finding common susceptibility variants for complex disease: past, present and future. *Brief Funct Genomic Proteomic* 2009; **8**: 345–352.
2 Ioannidis JP, Gwinn M, Little J: A road map for efficient and reliable human genome epidemiology. *Nat Genet* 2006; **38**: 3–5.
3 Hardy J, Singleton A: Genomewide association studies and human disease. *N Engl J Med* 2009; **360**: 1759–1768.
4 Khoury MJ, Bertram L, Boffetta P *et al*: Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. *Am J Epidemiol* 2009; **170**: 269–279.
5 Yu W, Clyne M, Dolan SM *et al*: GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC Bioinformatics* 2008; **9**: 205.
6 Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ: A navigator for human genome epidemiology. *Nat Genet* 2008; **40**: 124–125.
7 Dong LM, Potter JD, White E, Ulrich CM, Cardon LR, Peters U: Genetic susceptibility to cancer: the role of polymorphisms in candidate genes. *JAMA* 2008; **299**: 2423–2436.
8 Yu W, Clyne M, Khoury MJ, Gwinn M: Phenopedia and genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics* 2010; **26**: 145–146.