

A knowledge base for tracking the impact of genomics on population health

Wei Yu, MS, PhD¹, Marta Gwinn, MD, MPH^{1,2}, W. David Dotson, PhD¹,
Ridgely Fisk Green, MS, PhD^{1,3}, Mindy Clyne, MHS^{4,5}, Anja Wulf, BA^{1,6}, Scott Bowen, MPH¹,
Katherine Kolor, MS, PhD¹ and Muin J. Khoury, MD, PhD¹

Purpose: We created an online knowledge base (the Public Health Genomics Knowledge Base (PHGKB)) to provide systematically curated and updated information that bridges population-based research on genomics with clinical and public health applications.

Methods: Weekly horizon scanning of a wide variety of online resources is used to retrieve relevant scientific publications, guidelines, and commentaries. After curation by domain experts, links are deposited into Web-based databases.

Results: PHGKB currently consists of nine component databases. Users can search the entire knowledge base or search one or more component databases directly and choose options for customizing the display of their search results.

Conclusion: PHGKB offers researchers, policy makers, practitioners, and the general public a way to find information they need to understand the complicated landscape of genomics and population health.

Genet Med advance online publication 9 June 2016

INTRODUCTION

Genomic information is increasingly finding its way into clinical and public health practice, from diagnosis of rare genetic diseases to diagnosis and treatment of common chronic diseases and infections. The new Precision Medicine Initiative¹ and related developments will increase the number of people whose genomes are sequenced in the next decade. Meaningful clinical interpretation of emerging information requires the integration of data from basic, clinical, and population studies. Although such data are abundant, they are widely dispersed across the peer-reviewed literature and other online resources. Clinicians and public health professionals require credible information to use genomic information in practice.

In 2001, the Centers for Disease Control and Prevention (CDC) Office of Public Health Genomics began to systematically compile and curate an online catalog of published population-based studies of human gene–disease associations. In 2008, Office of Public Health Genomics launched the Human Genome Epidemiology Navigator (HuGE Navigator)² as an application for mining the rapidly growing database. By 2015, it contained citations for more than 100,000 scientific publications and had acquired more than 150,000 users (i.e., unique IP addresses). At that time, the scientific agenda and public interest were shifting increasingly from gene discovery to translation—that is, the use of genomic information to develop genome-based tests, drugs, and other applications.³ The Genomic Applications in Practice and Prevention Network was one of several government-sponsored, interdisciplinary efforts to address a perceived

“lack of readily accessible information about the utility of most genomic applications and the lack of necessary knowledge by consumers and providers to implement what is known.”⁴

To help address this need, we have taken what we view as the next step in showing how epidemiologic and other information can be used to improve population health: launching the Public Health Genomics Knowledge Base (PHGKB) (<http://phgkb.cdc.gov>). Our goal is to organize information from a wide variety of sources and in varying formats that are needed to describe the translational trajectories of genomic discoveries. Thus, although PHGKB's component databases have different formats and data structures and can be searched individually, searching PHGKB as a whole also produces seamless results. Here, we briefly describe PHGKB and present an initial cross-sectional analysis of its contents.

MATERIALS AND METHODS

PHGKB includes publications and other relevant Web-based resources captured by our weekly horizon scan. Some of our methods and early results have been described in previous publications.^{5,6} The content is indexed and grouped into categories that include practice guidelines, systematic reviews, implementation studies, and applications of genomic tests and family health history classified according to the level of available evidence.⁷ PHGKB was built using J2EE technology⁸ and other Java open-source frameworks, including Hibernate⁹ and Struts.¹⁰ As the largest constituent of PHGKB content, the scientific literature is represented by PubMed abstracts indexed with Medical Subject

¹Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, Georgia, USA; ²McKing Consulting Corporation, Atlanta, Georgia, USA; ³Carter Consulting, Inc., Atlanta, Georgia, USA; ⁴Epidemiology and Genomics Research Program, National Cancer Institute, Bethesda, Maryland, USA; ⁵Kelly Services, Troy, Michigan, USA; ⁶Cadence Group, Atlanta, Georgia, USA. Correspondence: Wei Yu (wby0@cdc.gov)

Submitted 12 February 2016; accepted 6 April 2016; advance online publication 9 June 2016. doi:10.1038/gim.2016.63

Headings (MeSH) terminology. Use of the MeSH tree hierarchies and the Unified Medical Language System metathesaurus enhances the system's search capacity.

RESULTS

PHGKB is an open access, Web-based, searchable database that provides access to a spectrum of information on genomics and population health, from basic research to implementation. PHGKB currently consists of nine component databases, including HuGE Navigator (Table 1). Users interested in specific topics can perform a global search of the entire knowledge base or search one or more component databases directly and choose options for customizing the display of their search results. The results of searching all databases in PHGKB are displayed according to steps in the translational pathway from discovery to implementation.¹¹ For example, results of a search for breast cancer (Figure 1) are arrayed from discovery to implementation, with special emphasis on evidence synthesis and guidelines, as well as on CDC products. Each category also includes links to specialized external resources.

PHGKB also offers users several ways to keep abreast of new information. First, the PHGKB main page features two sections that are updated almost daily: Hot Topics of the Day, curated by domain experts, and What's New, which displays recent additions to the database and summary statistics. In addition, two weekly e-mail newsletters are available by subscription—*Genomics & Health Impact Weekly Scan* and

Advanced Molecular Detection Clips (focused on human and pathogen genomics, respectively)—which direct users to new content posted on the PHGKB website.

DISCUSSION

Genomics has given rise to many specialized online databases that were designed primarily for use by researchers and other expert users. PHGKB is unique in providing systematically curated and updated information that bridges population-based research with clinical and public health applications. We acknowledge that PHGKB is not comprehensive, especially given the fluid state of translation and implementation research. Most genomic research is still focused on new discoveries; however, the focus of PHGKB is the small fraction—perhaps 1%—of genomics-related publications that address epidemiology, evaluation and evidence synthesis, implementation, and outcomes, as we have described elsewhere.⁵ Finding these needles in a haystack is important because they are most relevant to population health. As the knowledge base grows, it will become useful for tracing the translational trajectories of specific discoveries into clinical application and population health outcomes. So far, PHGKB has undergone limited pilot testing by selected users at the CDC and in state health departments. With this report, we invite potential users to explore the resource. We intend to conduct additional evaluation studies in the near future.

Genomic literacy is becoming a fundamental requirement for clinical and public health decision makers who have the power to improve patient and population health. We hope that PHGKB

Table 1 PHGKB component databases

Database name	Component database description	Number of records (as of 28 March 2016)
HuGE Literature Finder (HuGE Navigator)	Published population-based studies of genetic associations and other articles on human genome epidemiology	116,152
Hot Topics Database	Genomic insights into specific diseases and topics related to public health; sources include published scientific literature, reviews, blogs, and popular articles on genomic discoveries with potential applications to policy and practice	3,881
Genomics & Health Impact Scan Database	Published scientific literature on translation of genomics with potential to impact population health, focusing on noncommunicable diseases across the life span	4,718
Advanced Molecular Detection Clips Database	Published scientific literature, tools, and databases on the emerging role of pathogen genomics and host–pathogen interactions in infectious disease control and prevention	4,142
Guideline Database	Updated guidelines, policies, and recommendations on genomic research and practice, as provided by professional organizations, federal advisory groups, expert panels and policy groups	272
Evidence Classification Database	Currently available applications of genomic tests and family health history, classified into tiers 1–3 by level of evidence	159
Implementation Database	State and national activities that integrate genomics into public health programs and clinical practice; results can be filtered by state, condition, and resource type (data, programs, education, policy, tools, and general information)	347
CDC Information Database	General CDC public health information on specific diseases and health-related topics, including genomic information when available	610
CDC-Authored Genomics Publication Database	CDC-authored publications in public health genomics, including infectious diseases, newborn screening, reproductive health, genetic testing, cancer, chronic diseases, birth defects and developmental disabilities, environmental and occupational health, as well as laboratory, bioinformatics, and statistical methods	1,226

CDC, Centers for Disease Control and Prevention; PHGKB, Public Health Genomics Knowledge Base.

Public Health Genomics Knowledge Base (v1.0)

PHGKB
About
MyPHGKB
Genomics & Health Impact Weekly Scan (Current Edition)
Advanced Molecular Detection Weekly Clips (Current Edition)
Genomics (A-Z)
Office of Public Health Genomics
State Implementation Map
Relevant Resources
All Databases +
Contact Us

Search Knowledge Base
Enter a search term
Submit

Recommend Tweet Share

Search PHGKB: Search MyPHGKB

What's New

Last Updated: Apr 24, 2016

- Fine-scale mapping of 8q24 locus identifies multiple independent risk variants for breast cancer. [Sni Jiajun, et al. International journal of cancer 2016 4](#)
- Intensive breast screening in BRCA2 mutation carriers is associated with reduced breast cancer specific and all cause mortality. [Evans D G, et al. Hereditary cancer in clinical practice 2016 0 8](#)
- Characteristics, Malignancy Rate, and Follow-up of BI-RADS Category 3 Lesions Identified at Breast MR Imaging: Implications for MR Image Interpretation and Management. [Chikarmane Sona A, et al. Radiology 2016 4 151548](#)
- Frequency of germline DNA genetic findings in an unselected prospective cohort of triple-negative breast cancer patients participating in a platinum-based neoadjuvant chemotherapy trial. [González-Rivera Milagros, et al. Breast cancer research and treatment 2016 4 \(3\) 507-15](#)
- Incorporating truncating variants in PALB2, CHEK2, and ATM into the BOADICEA breast cancer risk model [Lee AJ, et al. Genetics in Medicine \(2016\) April 14](#)

More

CDC Resources

- CDC Information (9)
- CDC-Autored Pub (11)

Discoveries & Insights

- Human (67)
- Pathogen (1)

Epidemiology

- Human (5144)
- Pathogen (0)

Translational Research

- Human (402)
- Pathogen (0)

Evidence Synthesis

- Guidelines (24)
- Tier Table (19)
- Synthesis (49)

Practice & Implementation

- Implementation (116)

Figure 1 Results of a search performed on 24 April 2016 for breast cancer information using the Public Health Genomics Knowledge Base.

offers a useful resource to researchers, policy makers, practitioners, and members of the public who are interested in understanding how genomic research can contribute to better health.

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

1. National Institutes of Health. Precision Medicine Initiative Cohort Program. <https://www.nih.gov/precision-medicine-initiative-cohort-program>. Accessed 20 January 2016.
2. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. A navigator for human genome epidemiology. *Nat Genet* 2008;40:124–125.
3. Green ED, Guyer MS; National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* 2011;470:204–213.
4. Khoury MJ, Feero WG, Reyes M, et al.; GAPPNet Planning Group. The genomic applications in practice and prevention network. *Genet Med* 2009;11:488–494.
5. Clyne M, Schully SD, Dotson WD, et al. Horizon scanning for translational genomic research beyond bench to bedside. *Genet Med* 2014;16:535–538.
6. Yu W, Clyne M, Dolan SM, et al. GAPscreeener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC Bioinformatics* 2008;9:205.
7. Dotson WD, Douglas MP, Kolor K, et al. Prioritizing genomic applications for action by level of evidence: a horizon-scanning method. *Clin Pharmacol Ther* 2014;95:394–402.
8. ORACLE. Jave EE 7 SDK downloads. <http://www.oracle.com/technetwork/java/javase/downloads/index.html>. Accessed 20 December 2014.
9. Hibernate. <http://hibernate.org/>. Accessed 20 December 2014.
10. Apache Software Foundation. Apache Struts. <http://struts.apache.org/>. Accessed 20 December 2014.
11. Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L. 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* 2007;9:665–674.