European Journal of Human Genetics (2011) 19, 1095–1099

www.nature.com/ejhg

SHORT REPORT

GWAS Integrator: a bioinformatics tool to explore human genetic associations reported in published genome-wide association studies

Wei Yu*[,1], Ajay Yesupriya[1], Anja Wulf[1], Lucia A Hindorff[2], Nicole Dowling[1], Muin J Khoury[1] and Marta Gwinn[1]

Genome-wide association studies (GWAS) have successfully identified numerous genetic loci that are associated with phenotypic traits and diseases. GWAS Integrator is a bioinformatics tool that integrates information on these associations from the National Human Genome Research institute (NHGRI) Catalog, SNAP (SNP Annotation and Proxy Search), and the Human Genome Epidemiology (HuGE) Navigator literature database. This tool includes robust search and data mining functionalities that can be used to quickly identify relevant associations from GWAS, as well as proxy single-nucleotide polymorphisms (SNPs) and potential candidate genes. Query-based University of California Santa Cruz (UCSC) Genome Browser custom tracks are generated dynamically on the basis of users' selected GWAS hits or candidate genes from HuGE Navigator literature database (http://www.hugenavigator.net/HuGENavigator/gWAHitStartPage.do). The GWAS Integrator may help enhance inference on potential genetic associations identified from GWAS studies.
European Journal of Human Genetics (2011) **19,** 1095–1099; doi:10.1038/ejhg.2011.91; published online 25 May 2011

Keywords: genome-wide association studies; database; bioinformatics

INTRODUCTION

The completion of the human genome and HapMap projects combined with advances in high throughput genotyping techniques have resulted in an explosion of genome-wide association studies (GWAS).[1] These studies interrogate hundreds of thousands to a few million genetic variants and have identified a large number of loci associated with phenotypic traits or disease outcomes. As a result of their early and continued success, the number of published GWAS has steadily increased each year, from just two in 2005 to 238 in 2010 (as of 8 December; data from the statistic page in GWAS Integrator). To help the research community find these publications and further explore the reported associations, the National Human Genome Research institute (NHGRI) has established, and maintains the NHGRI GWAS Catalog (http://www.genome.gov/26525384), an online, regularly updated database of single nucleotide polymorphism (SNP)-trait associations from GWAS.[2] We have developed the GWAS Integrator, a bioinformatics tool that offers a robust search capacity and a set of data mining functions by integrating information from the NHGRI GWAS Catalog, with data from other established bioinformatics resources including HapMap (http://hapmap.ncbi.nlm.nih.gov/), the Human Genome Epidemiology (HuGE) Navigator (http://www.hugenavigator.net/), SNP Annotation and Proxy Search (SNAP) (http://www.broadinstitute.org/mpg/snap/ldsearch.php) and University of California Santa Cruz (UCSC) genome browser (http://genome.ucsc.edu/cgi-bin/hgGateway).

IMPLEMENTATION

The GWAS Integrator was built on J2EE technology (http://java.sun.com/javaee/) and on other Java open-source frameworks, including Hibernate (http://www.hibernate.org/), Strut (http://struts.apache.org/), and JChart (http://jcharts.krysalis.org/). The database is populated and updated with SNP-trait associations from the NHGRI GWAS Catalog each week when new associations are available; details about the selection criteria for these associations are available on the NHGRI GWAS Catalog website. Chromosomal locations of the associated SNPs and relevant proxy SNPs are downloaded from SNAP and UCSC, added to the database as needed (NCBI Build 36/UCSC Version 18 (hg18)). Records from the NCBI Entrez Gene database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gene) are used as standards for gene information, including chromosomal location. As a component of the HuGE Navigator,[3] the GWAS Integrator can take advantage of the established informatics infrastructure used in this integrated knowledge on the basis of human genome epidemiology. The HuGE literature database includes PubMed abstracts indexed with MeSH terminology (http://www.ncbi.nlm.nih.gov/sites/entrez?db=mesh); allowing use of the MeSH tree hierarchies and the Unified Medical Language System metathesaurus for mapping different synonyms of the phenotype/disease terms into a standard code enhances the search capacity of the GWAS Integrator. In addition, genes indexed in the HuGE literature database can be used to identify relevant candidate genes. The detailed schema for the HuGE Navigator database can be found in the paper by Yu *et al.*[4]

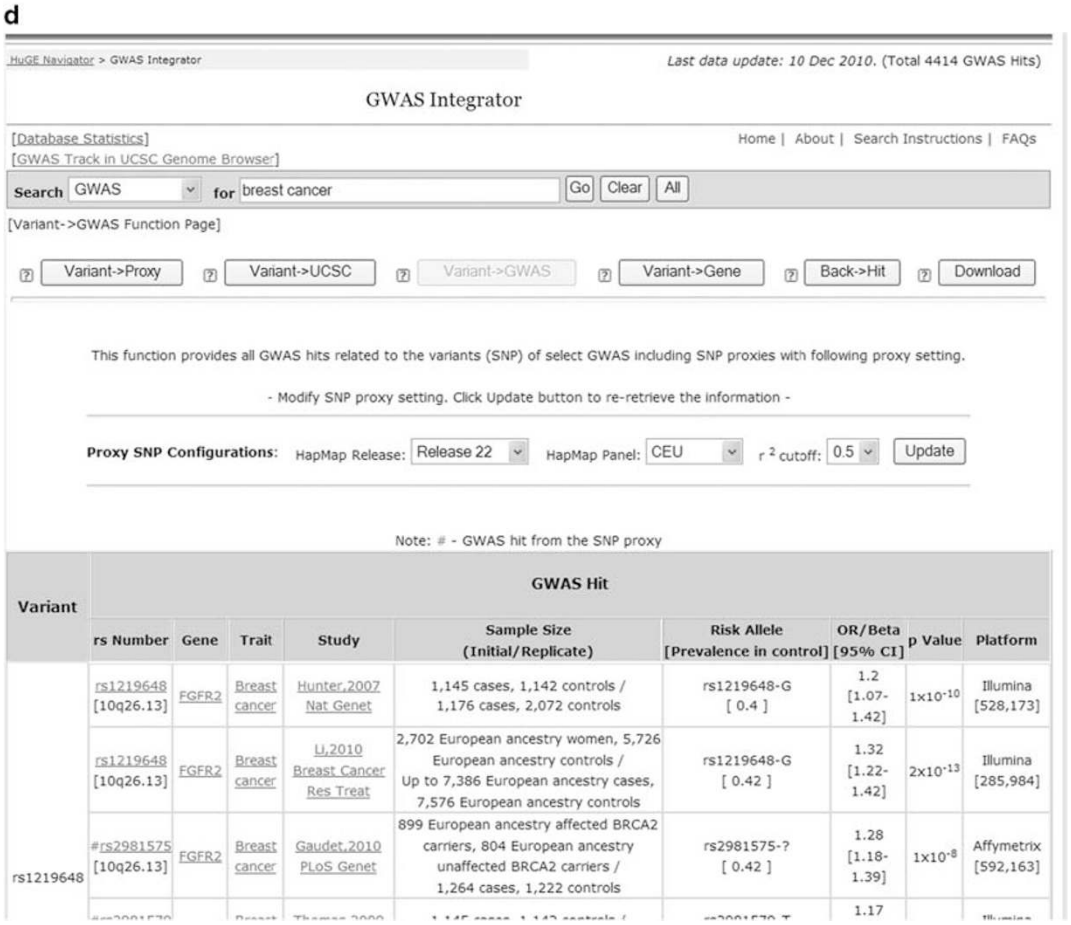
FEATURES

Robust search capacity

Users can perform free text searches of data extracted from published GWAS. Searchable terms include the disease/trait, gene name/gene

[1]Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, GA, USA; [2]Office of Population Genomics, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
*Correspondence: Dr W Yu, Office of Public Health Genomics, Centers for Disease Control and Prevention, 4770 Buford Highway, MS K-89, Atlanta, GA 30341, USA. Tel: +1 404 498 0053; Fax: +1 404 498 0140; E-mail: wby0@cdc.gov
Received 21 February 2011; revised 5 April 2011; accepted 21 April 2011; published online 25 May 2011

Figure 1 Continued.



Variant	GWAS Hit							
	rs Number	Gene	Trait	Study	Sample Size (Initial/Replicate)	Risk Allele [Prevalence in control]	OR/Beta [95% CI]	p Value	Platform
rs1219648	rs1219648 [10q26.13]	FGFR2	Breast cancer	Hunter,2007 Nat Genet	1,145 cases, 1,142 controls / 1,176 cases, 2,072 controls	rs1219648-G [0.4]	1.2 [1.07-1.42]	$1x10^{-10}$	Illumina [528,173]
	rs1219648 [10q26.13]	FGFR2	Breast cancer	U,2010 Breast Cancer Res Treat	2,702 European ancestry women, 5,726 European ancestry controls / Up to 7,386 European ancestry cases, 7,576 European ancestry controls	rs1219648-G [0.42]	1.32 [1.22-1.42]	$2x10^{-13}$	Illumina [285,984]
	#rs2981575 [10q26.13]	FGFR2	Breast cancer	Gaudet,2010 PLoS Genet	899 European ancestry affected BRCA2 carriers, 804 European ancestry unaffected BRCA2 carriers / 1,264 cases, 1,222 controls	rs2981575-? [0.42]	1.28 [1.18-1.39]	$1x10^{-8}$	Affymetrix [592,163]
							1.17		

Figure 1 Continued.

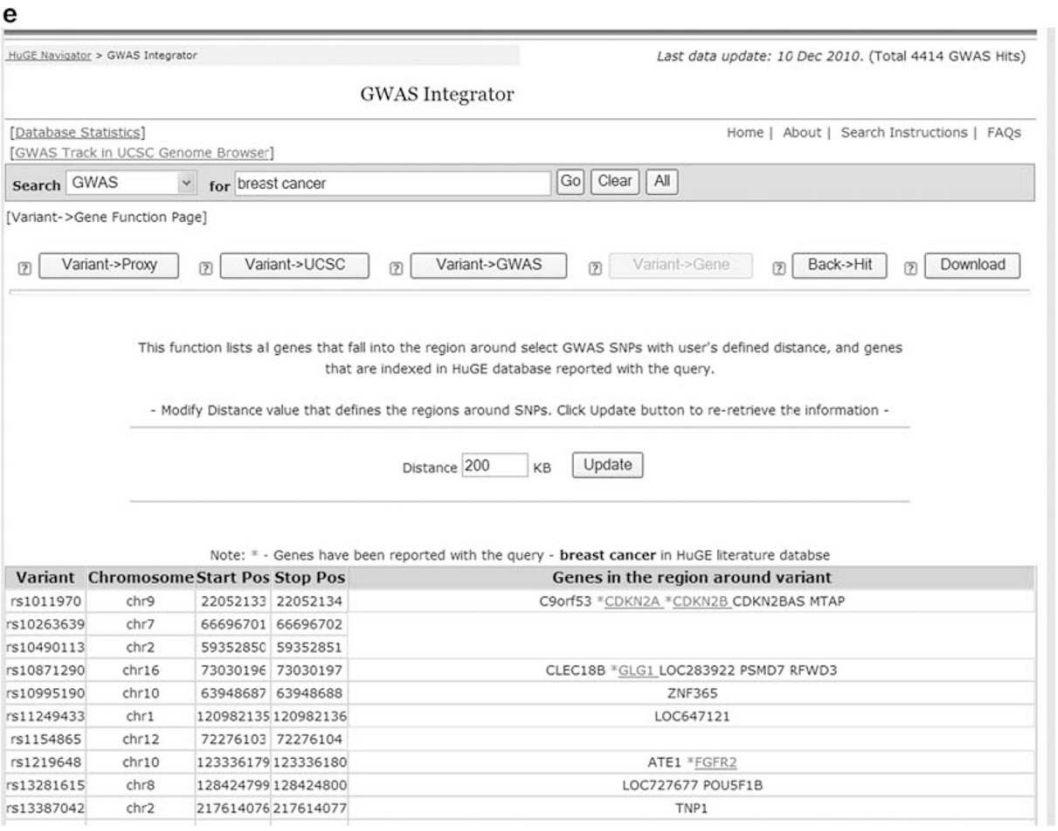


Figure 1 Illustration screen shots for GWAS Integrator. (**a**) Display of the GWAS hits related to breast cancer. (**b**) Display of SNP proxies of the variants related breast cancer. (**c**) Display of dynamically-generated SNP and candidate gene UCSC custom tracks related to breast cancer. (**d**) Display of GWAS hits from proxy SNPs of GWAS hits related to breast cancer. (**e**) Display of all genes that fall into the region around the selected GWAS hits related to breast cancer.

symbol/gene alias, rs number, first author name, journal, chromosome region, platform, PubMed ID, and any text in the publication title or abstract. Search results can be filtered by variant, gene, region, trait, publication, author, journal, and year, as well as by 'hit' (ie, the SNP-trait association identified in a GWAS). Results can be filtered multiple times. The filtering function also can be used to obtain a quick snapshot of GWAS published in a particular research field. For example, a user can easily get descriptive statistics for GWAS on breast cancer, including the number of variants that have been studied, the number of GWAS publications, etc (Figure 1a).

Data mining capacity

A series of data mining capacities can be used to further explore search results.

Variant->proxy function. This function provides information on SNP proxies related to the variants (SNP) of the selected GWAS hits. Users can define configuration parameters for proxy SNP retrieval, such as the HapMap release version, HapMap population, and r2 cutoff (Figure 1b).

Variant->UCSC function. This function dynamically creates an SNP custom track to display selected GWAS hits in the UCSC Genome Browser. Users can select the SNP to center the display in the UCSC Genome Browser using a dropdown menu, which lists all the rs numbers for the selected GWAS hits. The 'Window Size' field defines the display range around the centered SNP in the UCSC Genome Browser; for example, when 500 kb is specified in the Window Size field, the UCSC Genome Browser will display 250 kb on each side of the centered SNP. Users can also include proxy SNPs in the SNP custom track, or create a separate custom track for genes indexed in the HuGE literature database related to the query (Figure 1c).

Variant->GWAS function. This function uses proxy SNPs to identify additional GWAS hits that may be related to the user-selected GWAS hits. Users can define configuration parameters for proxy SNP retrieval, such as HapMap release version, HapMap population, and r2 cutoff (Figure 1d).

Variant->gene function. This function lists all genes that fall into the region around the selected GWAS Hits. Users can define the genomic distance around the hits. Genes that are also indexed in the HuGE literature database and reported with the query term are highlighted with a hyperlink to the corresponding Genopedia[5] record in HuGE Navigator (Figure 1e).

Proxy reference search

Users can also search for variant-trait associations using proxy SNPs. For example, searching with 'rs663129' will lead to six proxy SNPs that have GWAS hits.

Real-time tracking

The statistics page presents an overview of published GWAS, including total numbers of publications, hits, reported genes, genic SNPs,

intergenic SNPs, variants, and disease/traits. Temporal trends are displayed graphically for each item. A top 10 list is generated and displayed in web tables, including variant, gene, chromosome region, disease/trait, first author, and journal. As of 10 February 2011, the database contains 4817 GWAS hits, representing 475 disease/traits and 3920 variants from 796 publications.

CONCLUSION

GWAS of phenotypic traits and diseases have successfully identified a large number of genetic loci for further investigation by replication, meta-analysis, imputation of untyped loci, resequencing, identification of functional polymorphisms, and analysis of gene–gene and gene–environment interactions.[6] By integrating relevant information from multiple data sources, the GWAS Integrator helps researchers to quickly identify GWAS of interest, examine the findings in the context of other genetic and epidemiologic research, perform on-line data mining, and make inferences that can inform future studies. Including the GWAS Integrator in the HuGE Navigator allows it to take advantage of the established informatics infrastructure of one of the most comprehensive repositories of published genetic associations – the HuGE literature database. The dynamic generation of a custom track is an efficient way to access all features offered by the UCSC genome browser. Ongoing collaboration between CDC and NHGRI in collecting and synchronizing GWAS data will guarantee the most updated GWAS data sources. As a new application in HuGE Navigator, GWAS Integrator was built to interconnect to other applications already in the system, such as Genopedia, Gene Prospector, etc, so that navigation to other information is provided. Although GWAS Integrator database content is currently limited to the NHGRI GWAS Catalog, we plan to implement a feature that allows users to import their own GWAS data for data mining.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

1 Hardy J, Singleton A: Genomewide association studies and human disease. *N Engl J Med* 2009; **360**: 1759–1768.
2 Hindorff LA, Sethupathy P, Junkins HA *et al*: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009; **106**: 9362–9367.
3 Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ: A navigator for human genome epidemiology. *Nat Genet* 2008; **40**: 124–125.
4 Yu W, Yesupriya A, Wulf A, Qu J, Khoury MJ, Gwinn M: 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* 2007; **8**: 436.
5 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.
6 Panoutsopoulou K, Zeggini E: Finding common susceptibility variants for complex disease: past, present and future. *Brief Funct Genomic Proteomic* 2009; **8**: 345–352.