European Journal of Human Genetics (2013) 21, 471–473

www.nature.com/ejhg

SHORT REPORT

WikiGWA: an open platform for collecting and using genome-wide association results

Jie Huang*[,1], Eric Y Liu[2], Ryan Welch[3], Cristen Willer[4], Lucia A Hindorff[5] and Yun Li*[,2,6]

The number of discovered genetic variants from genome-wide association (GWA) studies (GWAS) has been growing rapidly. Centralized efforts such as the National Human Genome Research Institute's GWAS catalog provide regular updates and a convenient interface for quick lookup. However, the catalog entries are manually curated and rely on data from published articles. Other tools such as SNPedia (http://www.snpedia.com) collect published results regarding functional consequences of genetic variations. Here, we propose an approach that allows individual investigators to share their GWA results through an open platform. Unlike GWAS catalog or SNPedia, wikiGWA collects first-hand GWAS results and in a much larger scale. Investigators are not only able to post a much larger amount of results, but also post results from unpublished studies, which could alleviate publication bias and facilitate identification of weak signals. Our interface allows for flexible and fast queries, and the query results are formatted to work seamlessly with the LocusZoom program for visualization and annotation. We here describe wikiGWA, made publically available at http://www.wikiGWA.org.
European Journal of Human Genetics (2013) **21,** 471–473; doi:10.1038/ejhg.2012.187; published online 29 August 2012

Keywords: genome-wide association; open platform; bioinformatics

INTRODUCTION

The proliferation of genome-wide association study (GWAS) findings has encouraged the development of resources to browse and use these data. The National Human Genome Research Institute's GWAS catalog paper[1] has been formally cited >600 times, revealing the great impact of a publically available GWA results collection on the broad scientific community. As a manually curated resource, this catalog extracts a limited number of SNPs from each published GWA study, prioritizing SNPs replicated for association with phenotypic trait(s) (methods available at http://www.genome.gov/gwastudies). A second National Institute of Health (NIH) database of GWA findings[2] collects SNP-phenotype association results with a less stringent significance threshold. There are also other initiatives and technical applications within NIH that integrate GWAS data with various genomic databases, including the Phenotype–Genotype Integrator (PheGenI: http://www.ncbi.nlm.nih.gov/gap/PheGenI) and the Center for Disease Control and Prevention's GWAS Integrator (http://www.hugenavigator.net/HuGENavigator/gWAHitStartPage.do). These resources, based on peer-reviewed findings and curated by teams within national governmental research agencies, provide trustworthy information and highly regarded service to the broad research community.

Among the non-NIH resources, GWAS Central has an objective close to that of our platform, which is to 'actively gather datasets from public domain projects, and encourage direct data submission from the community' (http://www.gwascentral.org/). However, we found the query tool is returning limited results and the submission of large data from users seems not readily assessable, with an Excel file data submission template that has hundreds of variables. Other 'grassroots' tools such as SNPedia provide a more collaborative and open approach for collecting and sharing information about genetic variations.[3] However, that body of information is for a quite different purpose and audience than what we are proposing here: SNPedia is a wiki resource regarding the functional consequences of human genetic variation based on published studies, whereas our wikiGWA program intends to collect first-hand GWA results from and for genetic researchers. Our wikiGWA platform adopts the wiki philosophy where content comes from a broad community and is then shared and utilized. However, all the upfront interface and backend databases are developed by specialized programmers instead of relying on the publically available Wiki developing programs, therefore giving us more flexibility to design a platform that meets user's needs.

All other existing programs and efforts mentioned above work from published results. However, for our wikiGWA platform, although we still expect the majority of the user input data to come from published or pre-published studies, we certainly do not prevent researchers from sharing their unpublished results, thus would effectively alleviate publication bias and provide an invaluable resource for researchers interested in first-hand findings. Even for the same GWAS, researchers will be able to submit results based on different models of inheritance, stratified by factors of interest, etc. Thus, this platform will facilitate dissemination of GWAS data by

[1]Department of Human Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK; [2]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [3]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA; [4]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA; [5]Office of Population Genomics, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA; [6]Department of Genetics, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
*Correspondence: Dr J Huang, Department of Human Genetics, Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge CB10 1SA, UK.
Tel: 44 (0) 1223 498635; Fax: 44 (0) 1223 491919; E-mail: jh21@sanger.ac.uk
or Dr Y Li, Department of Genetics and Biostatistics, University of North Carolina, Room 5090 GMB, 120 Mason Farm Road, CB7264, Chapel Hill, NC 27599-7264, USA.
Tel: +1 919 843 2832; Fax: +1 919 843 4682; E-mail: yunli@med.unc.edu
Received 16 April 2012; revised 18 July 2012; accepted 18 July 2012; published online 29 August 2012

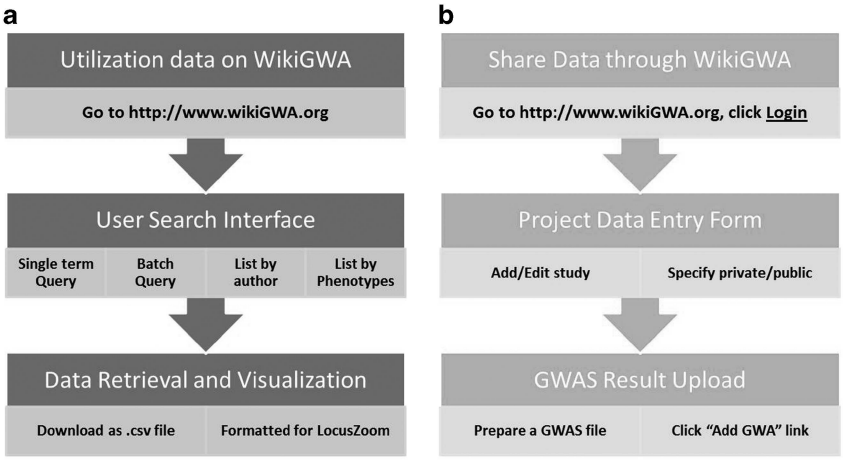

Figure 1 The design and flow of wikiGWA (**a**) for utilizing data in wikiGWA; (**b**) for sharing data through wikiGWA.

allowing sharing both publicly and privately and by collecting structured data that in a form complementary to other GWAS resources.

MATERIALS AND METHODS

Here, we propose a Wikipedia style platform for researchers to share their GWA findings. We designed and developed the backend databases and the upfront user interface to provide the following three core features:

1. User interface for GWA data collection: Anyone may upload GWA results, at any *P*-value threshold considered appropriate. Users enter study level information first and then upload GWA results for each study. For studies already entered into the platform, users can edit any study level data, and selectively delete and edit GWA results for each study at their discretion. To reduce publication bias, we enable and encourage the sharing of unpublished results. For uploading unpublished data, some fields including publication date, journal, and authors are optional. Up to 1 million SNP association data could be uploaded in one single upload at this moment, which could be increased when needed.
2. User interface for GWA queries: We design a user-friendly query interface to allow searching on any combination of fields stored, including sample size, publication date, and ethnicity. Furthermore, we allow both single term query and batch query. The batch query facilitates the lookup of multiple SNPs, for example, the top 1000 SNPs from the user's current GWA study. A user can view the search results sorted by any field. A user can also choose to download all search results in a spreadsheet. We optimize our database design and server-side program to allow fast query even with a large amount of SNPs stored. Investigators who do not want to share data can still upload data by specifying it as private and then using the downstream annotation and plotting features provided by wikiGWA.
3. Annotation and visualization for regions of interest: WikiGWA allows users to specify the inclusion of SNPs within certain proximity of their interested SNPs (with 100 kb, for example) so that the residing genomic regions can be examined in greater detail. The output data from a query are downloadable in the input format of LocusZoom[4] for SNP annotations and regional visualization. We customized the LocusZoom program to display a PRIMe[5] style legend, where SNPs are annotated by their associated studies when SNPs from multiple studies are plotted together. Each individual SNP record is also linked to the UCSC Genome Browser to display a more interactive and detailed visualization.

RESULTS

The main result of our open platform is the creation of the website http://www.wikiGWA.org and its associated databases.

The two panels in Figure 1 shows the flow chart for utilization and posting GWA data through our platform, respectively. Beside the upfront user interface and the underlying databases, we also implemented scripts to perform sanity checks on data uploaded to wikiGWA. For example, users will be notified by email after they uploaded data that contains negative values for odds ratios. For published studies, the PMID uniqueness is checked and enforced to avoid uploading the same GWA study multiple times. For unpublished or pre-published study, a warning message will be displayed when a user adds a study whose cohort and phenotype are already included in the platform. We will keep monitoring user-uploaded data flowing into our platform and imposing more comprehensive sanity checks along the way.

By the time of the submission of this manuscript, $>300\,000$ SNP associations have been entered into the platform, a majority of which are in the cardiovascular and metabolic phenotypes category. We duplicated this data 100 times to create a repository of 30 million records to test the scalability of wikiGWA. We found the site operates without noticeable delay or malfunction.

Figure 2 shows the LocusZoom plot with data coming directly from wikiGWA by querying the region on chromosome 11 between 42.2 and 42.4 Mb region. This provides a straightforward and intuitive channel for researchers who are interested in potential pleiotropy in this region.

DISCUSSION

In summary, we present here an open platform for investigators to share and use first-hand GWAS findings. Our platform allows sharing of large amount of GWA results even for unpublished studies. It also provides a user friendly interface to quickly query data of interest from a large GWAS results collection. It has been tested to hold phenotype–genotype associations for tens of millions of records. Finally, it works seamlessly with a widely used plotting and annotation tool, LocusZoom.

It was reported that using a combination of personal genotype and summary results of GWA could potentially reveal the identity of GWA participants.[6] However, this theoretic risk has been explained and played down in practical situations by those who are aiming to collect dense GWA results and those who are striving for a balance research with privacy and protection.[7,8] Nevertheless, to better protect privacy, we currently enforce a policy where a maximum number of 1000 SNP results could be returned from a user query.

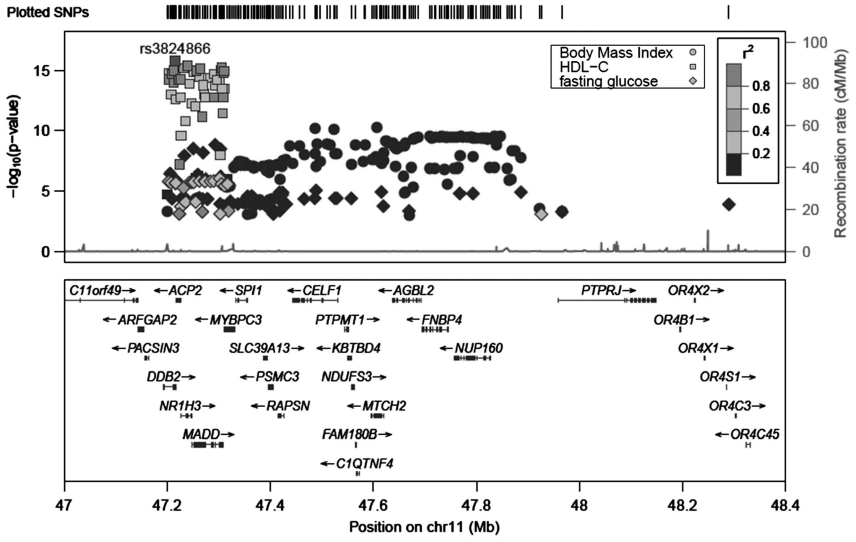

Figure 2 A wikiGWA query result displayed in LocusZoom.

To ensure data accuracy and integrity, our program takes an all-or-none approach for uploading and editing of GWA results. That is, a user can batch upload a full or partial GWA results file and have the option to delete the whole uploaded data of a GWA study, but the user cannot edit individual SNPs. The uploaded file format includes fields for both SNP names and genomic positions. New data entered through our platform are automatically validated for genomic positions, by using map files compiled from dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/).

Although our initial target is to share GWA results, we will not disable or discourage the sharing of results of similar format from candidate gene studies. We also encourage unpublished and pre-published data. The majority of the existing data comes from a GWAS study using 1000 Genomes based imputation.[9] Albeit grassroots, our design and development team (all six coauthors of this project) will monitor closely the functionality and efficiency of wikiGWA. We have already received users' feedback through email and the issues were directed to our development team and addressed accordingly. This provides a good model for us to work closely and collaboratively with a large research community to improve this platform and therefore the sharing and utilizing of large-scale GWA findings.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank the Consortium of GIANT and MAGIC for making a few their GWA results publically available. We thank the users who signed up and uploaded their GWA findings to our beta version website. YL is supported by the NIH grant R01-HG006292, R01-HG006703-01, and 3-R01-CA082659-11S1.

1 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.
2 Johnson AD, O'Donnell CJ: An open access database of genome-wide association results. *BMC Med Genet* 2009; **10**: 6.
3 Cariaso M, Lennon G: SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res* 2012; **40**: D1308–D1312.
4 Pruim RJ, Welch RP, Sanna S *et al*: LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010; **26**: 2336–2337.
5 Huang J, Johnson AD, O'Donnell CJ: PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies. *Bioinformatics* 2011; **27**: 1201–1206.
6 Homer N, Szelinger S, Redman M *et al*: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 2008; **4**: e1000167.
7 Church G, Heeney C, Hawkins N *et al*: Public access to genome-wide data: five views on balancing research with privacy and protection. *PLoS Genet* 2009; **5**: e1000665.
8 Johnson AD, Leslie R, O'Donnell CJ: Temporal trends in results availability from genome-wide association studies. *PLoS Genet* 2011; **7**: e1002269.
9 Huang J, Ellinghaus D, Franke A, Howie B, Li Y: 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *Eur J Hum Genet* 2012; **20**: 801–805.