Published online 1 December 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D1047–D1054***
doi:10.1093/nar/gkr1182

GWASdb: a database for human genetic variants identified by genome-wide association studies

Mulin Jun Li[1], Panwen Wang[1], Xiaorong Liu[1], Ee Lyn Lim[1,2], Zhangyong Wang[1,3], Meredith Yeager[4,5], Maria P. Wong[6], Pak Chung Sham[7], Stephen J. Chanock[5] and Junwen Wang[1,*]

[1]Department of Biochemistry, The University of Hong Kong, Hong Kong SAR, China, [2]Trinity College, University of Oxford, Oxford, UK, [3]Department of Computer Science, UCLA, Los Angeles, CA, [4]Core Genotyping Facility, SAIC-Frederick Inc., Frederick, MD, USA, [5]Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD, USA, [6]Department of Pathology and [7]Department of Psychiatry, The University of Hong Kong, Hong Kong SAR, China

Received August 11, 2011; Revised October 11, 2011; Accepted November 14, 2011

ABSTRACT

Recent advances in genome-wide association studies (GWAS) have enabled us to identify thousands of genetic variants (GVs) that are associated with human diseases. As next-generation sequencing technologies become less expensive, more GVs will be discovered in the near future. Existing databases, such as NHGRI GWAS Catalog, collect GVs with only genome-wide level significance. However, many true disease susceptibility loci have relatively moderate *P* values and are not included in these databases. We have developed GWASdb that contains 20 times more data than the GWAS Catalog and includes less significant GVs ($P < 1.0 \times 10^{-3}$) manually curated from the literature. In addition, GWASdb provides comprehensive functional annotations for each GV, including genomic mapping information, regulatory effects (transcription factor binding sites, microRNA target sites and splicing sites), amino acid substitutions, evolution, gene expression and disease associations. Furthermore, GWASdb classifies these GVs according to diseases using Disease-Ontology Lite and Human Phenotype Ontology. It can conduct pathway enrichment and PPI network association analysis for these diseases. GWASdb provides an intuitive, multifunctional database for biologists and clinicians to explore GVs and their functional inferences. It is freely available at http://jjwanglab.org/gwasdb and will be updated frequently.

INTRODUCTION

Thousands of genetic variants (GVs) associated with human traits and diseases have been identified by genome-wide association studies (GWAS). The advent of high throughput technologies, such as next-generation sequencing and very high-density microarrays, enable us to capture genome-wide variation on a much larger scale. With increasing sample sizes, GWAS studies based on these technologies will produce more information at higher resolutions. We will be able to detect many traits/diseases associated GVs, such as single nucleotide polymorphisms (SNPs), copy number variations (CNVs), and insertions and deletions (Indels) (1,2).

To understand the underlying regulatory and metabolic significance of these GVs, we have to consider biological evidences from different sources. However, in developing databases and web resources to integrate multidimensional functional annotations, researchers will inevitably encounter the following difficulties: (i) Searching and gathering GWAS results from published data for a specific trait/disease can be tedious and time-consuming. Researchers have to locate the publications by searching PubMed or other databases, and then gather GVs by manual curation either from the main text or from related supplementary materials for each publication. (ii) Individual curation lacks a universal criterion for data handling, which might cause data inconsistency and consequently affects the quality of the downstream analysis. (iii) Inference of the functional role of these GVs from heterogeneous databases will also be a challenge. Information (genomic elements, genetic and disease associated attributes) from different databases (such as dbSNP, HapMap, RefSeq, Ensemble and

*To whom correspondence should be addressed. Tel: +852 2819 2809; Fax: +852 2855 1254; Email: junwen@uw.edu

OMIM) needs to be gathered. If the information is not readily available, functional prediction will need to be performed using various available software or web servers.

Fortunately, several databases and tools have been developed to cope with these problems. The NHGRI GWAS Catalog has collected more than 5800 GVs from published GWAS (up to August 2011). The database used GWAS studies reporting at least one GV with $P < 5.0 \times 10^{-8}$, and the collected GVs were limited to $P < 1.0 \times 10^{-5}$ (3). This database also contains some statistical features including odds ratios and estimated risk allele frequencies. Johnson and O'Donnell have published a full gene-annotated GWAS database, which contains 56 411 GWAS genotype–phenotype associations with a threshold $P < 1.0 \times 10^{-3}$ (4). GWAS Central (previously named HGVbaseG2P) is another manually curated database that provides a centralized compilation of high level summary data from genetic association studies (5). Other databases also focus on data integration of GVs from GWAS, such as dbGaP PheGenI (6), Genetic Association Database (GAD) (7), HuGE Navigator (8), Varietas (9) and Snpedia. Many bioinformatics tools have been developed to quickly locate genome elements around GVs and to infer their putative functions, such as SNPselector (10), SNP Function Portal (11), F-SNP (12), SNPit (13), SNPLogic (14), SNPnexus (15), SCAN (16), GWAS analyzer (17) and pfSNP (18). These Web-based resources continually strive to provide a comprehensive knowledge base of the characteristics and functions of GVs.

However, existing resources also have limitations in satisfying the increasing demands of current GWAS research: (i) Many true disease susceptibility loci have relatively moderate P values which are ignored in existing databases. GVs with moderate effect sizes, usually filtered by strict cutoffs, can be directly related to diseases through gene–gene interaction in the context of regulatory networks and pathways (19). (ii) Most of the existing databases focus only on one or several aspects of the functional annotations, and not on GV-disease relationships. An integrative, comprehensive, up-to-date GWAS-based knowledge base that focuses on disease classification is needed.

Here, we present GWASdb, a user friendly database that combines collections of GVs from GWAS together with their functional annotations and disease classifications. We aim to provide an integrative, multidimensional functional annotation portal to help researchers and clinicians maximize the usage of the most recent GWAS data. The database provides the following information: (i) In addition to all the GVs annotated in the NHGRI GWAS Catalog, we manually curated the GVs that are marginally significant ($P < 1.0 \times 10^{-3}$) collected from supplementary materials of each original publication. (ii) We provide extensive functional annotations for these GVs. (iii) The GVs have been manually classified according to disease using Disease-Ontology Lite (DOLite) and Human Phenotype Ontology (HPO). The database can be used to conduct gene-based pathway enrichment and PPI network association analysis for diseases with sufficient variants.

DATABASE DESIGN AND CONTENT

We provide an intuitive, well-organized and easy-to-use web interface that allows users to explore the GVs from different perspectives, including genome, disease, gene regulation and protein interactions. Users can quickly search and locate a queried region by inputting the dbSNP id, gene symbol and chromosome region, or by directly clicking the data point on the plot of the GWAS overview. We have also developed a web-based genome viewer (Gviewer) to dynamically display the related information. Furthermore, to facilitate communication with other servers, we provide web service interfaces for machine-based large-scale data retrieval. We anticipate the database can facilitate follow-up analysis of specific diseases and can help researchers generate hypotheses by integrating multidimensional information concerning the target GV. The overall structure of our database is shown in Figure 1.

Data curation and collection

One major source of data for GWASdb was from the NHGRI GWAS Catalog. The GWAS Catalog has collected data on thousands of GVs from the literature, adopting a stringent criterion to ensure data consistency and integrity. SNP-trait associations for each GV gathered from each paper were limited to $P < 1.0 \times 10^{-5}$, and the database also restricted the number of SNP-trait associations extracted from each paper to 50 (3). We extended the scope of this database by using a relatively loose cutoff of $P < 1.0 \times 10^{-3}$ for data from each paper, and where possible GVs were included from supplementary materials. We used the same standards for other criteria, including P values derived from the largest sample size, population selected from a combined analysis or the largest one. Our purpose was to incorporate more GVs with moderate P values and to have more comprehensive functional annotations. At this current stage, we have gathered 70 411 GVs, 64 000 more than in the NHGRI GWAS Catalog (see Supplementary Data). Other well-organized GWAS databases also incorporated were Johnson and O'Donnell (4), dbGaP PheGenI (6), GAD (7), GWASCentral (5) and PharmGKB (20). We found many overlapping GVs annotated in these databases, which we combined by selecting only the most significant ones from the redundant GVs. We also omitted the GVs that we had already included from the NHGRI GWAS Catalog. In total, we obtained 146 537 GVs from the consolidation of several databases, 20 times more than in the NHGRI GWAS Catalog (see Supplementary Data). All the GVs can be viewed at either the whole genome level or at the chromosome level using the circular genome plot.

Constructing GV functional annotation

All the collected GVs were mapped to the latest dbSNP132 database. We then integrated comprehensive annotations from various sources for these GVs. These annotations were systematically divided into seven categories as follows: GV summary, genomic mapping, regulatory effect, amino acid substitution, evolution,

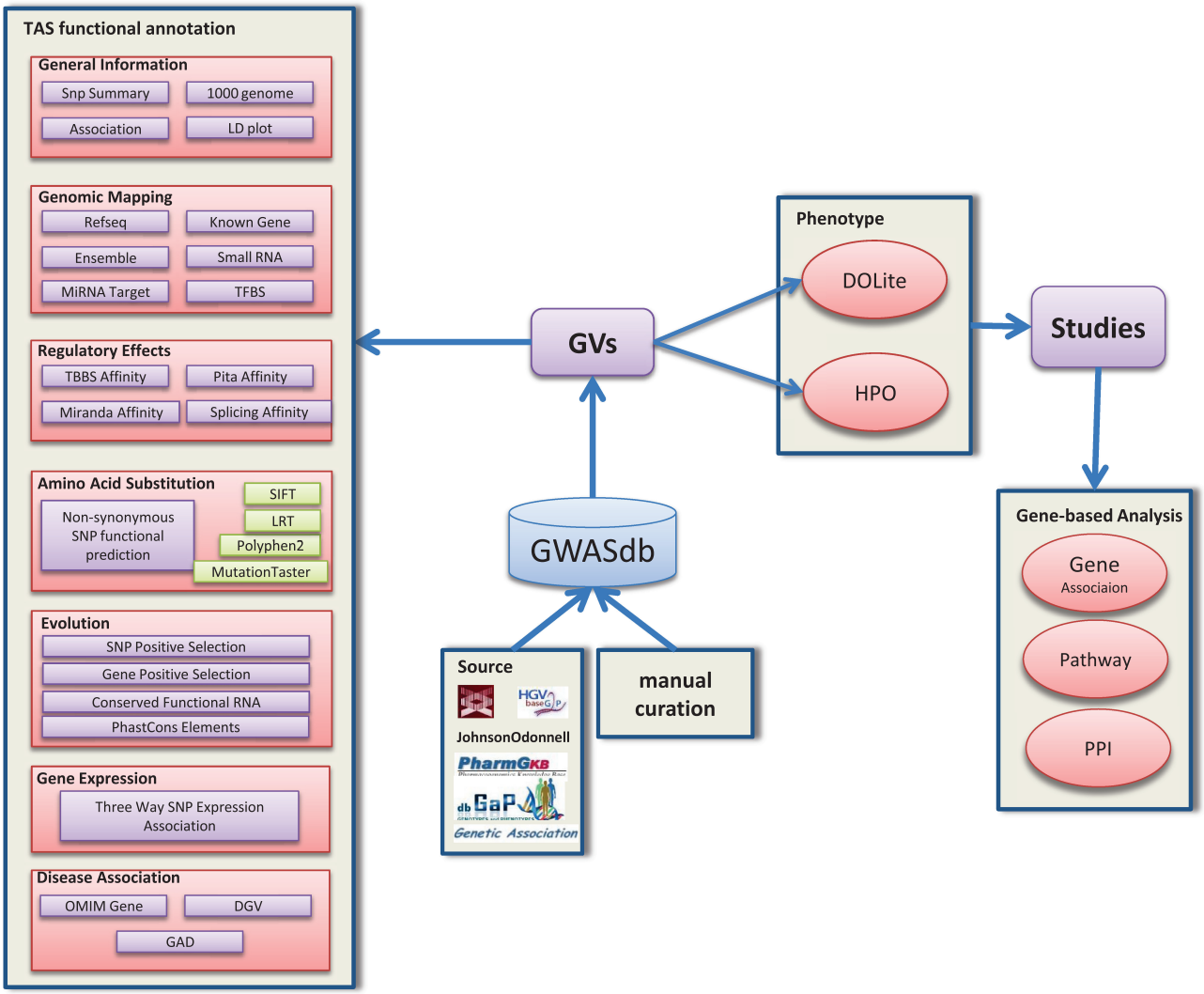

Figure 1. The overview of GWASdb database design. GWASdb consists of three main functions: precise scientific curation and resources integration on GWAS, comprehensive annotation of genetic variants and disease-oriented analysis in terms of DOLite and HPO.

gene expression and disease annotation (Table 1). For each category, we investigated the possible functional roles of each selected GV. For example, in the category of regulatory effect, we computed the affinity changes caused by different alleles of each GV, such as the affinities between transcriptional factors and their binding sites (21–23), microRNAs and their targets (23), and predicted splicing sites (18). The statistical significances of the binding affinity changes were calculated based on permutations of the binding partners (24).

We further calculated how the annotated GVs are distributed in different genomic regions. As shown in Figure 2a, 43.5% of all GVs are in the gene regions, such as intron, nonsense, missense, cds-indel, cds-synon, frameshift, 3′-UTR, 5′-UTR, 3′-nearGene and 5′-nearGene, as defined by dbSNP132. The rest of the GVs (~56.5%) are located in intergenic regions, which are areas that contain enhancers, promoter elements and many other long range regulators, and thus may be involved in gene regulation and regulatory networks (25). The top 15 traits/diseases with the most abundant GVs in our database are shown in Figure 2b.

Mapping of GVs using DOLite and HPO

DOLite is a simplified annotation of gene–disease associations. It was constructed from the OBO Foundry Disease Ontology (26). DOLite uses 561 independent nodes to describe gene–disease associations and is highly suited for GV-disease mapping in our database. We were able to successfully map 70% of our GVs into these nodes. However, DOLite does not include other phenotypes, such as height, weight and addiction, so another ontology database, HPO (27), was used. We were able to successfully map the rest of the GVs in our database in terms of HPO.

Disease-oriented analysis using DOLite and HPO

The mapping of GVs to diseases enables us to perform disease level meta-analysis. It is important to understand

Table 1. Description of annotations organized in GWASdb

Level	Item	Description	Reference
Snp Summary	General information	dbSNP 132 annotation for each GV	dbSNP-Q (32)
	Genome-wide association	Manual curation and collection	GWASdb
	1000 Genome SNP	SNPs and indels in 1000 Genomes Project 1049 subjects (May 2011 release)	1000 genome project
	LD plot	LD data from HapMap Phase II+III	HapMap
Genomic mapping	Reference gene	Gene annotation from NCBI Refseq	NCBI Refseq
	Ensemble gene	Gene annotation from Ensemble	Ensemble
	Known gene	Gene annotation from UCSC	UCSC
	Small RNA	snoRNA and miRNA annotations from UCSC	UCSC
	MicroRNA target	TargetScan generated miRNA target site predictions	UCSC
	Transcriptional factor binding site	Transcription factor binding sites conserved in the human/mouse/rat alignment, based on transfac Matrix Database (v7.0)	UCSC
	Enhancer	Human Enhancer verified by experiment	VISTA Enhancer DB (33)
	Insulator	CTCF binding site database for characterization of human genomic insulators	CTCFBSDB (34)
Regulatory effects	Transcriptional factor binding site affinity	GV affinity of TFBS prediction based on fold energy change with PWM scanning	GWASdb, TRANSFAC (35) JASPAR (36), UniPROBE (37)
	MicroRNA target site affinity (for Pita)	GV affinity of miRNA target prediction based on fold and hybrid energy change for PITA top targets	GWASdb, PITA (38)
	MicroRNA target site affinity (for Miranda)	GV affinity of miRNA target prediction based on hybrid energy change for miRanda targets	GWASdb, miRanda (39)
	Splicing site affinity	GV affinity of splicing site prediction	ssSNPTarget (40)
Amino acid substitution	Non-synonymous SNP functional prediction	Non-synonymous GV deterioration prediction	dbNSFP (41)
Evolution	SNP positive selection	The estimation of FST and heterozygosity of GV for positive selection	SNP@Evolution (42)
	Gene positive selection	The estimation of FST and heterozygosity of gene for positive selection	SNP@Evolution
	Conserved functional RNA	Conserved functional RNA, through RNA secondary structure predictions made with the EvoFold program	UCSC
	Conserved elements	Conserved elements produced by the PhastCons program based on a whole-genome alignment of vertebrates	UCSC
Gene expression	Three way SNP expression association	Gene co-expression relationships with GV effect	SNPxGE2 (43)
Disease association	OMIM	Online Mendelian Inheritance in Man	OMIM
	DGV	Curated catalogue of structural variation in the human genome	Database of Genome Variants
	GAD	Archive of human genetic association studies of complex diseases and disorders	Genetic Association Database

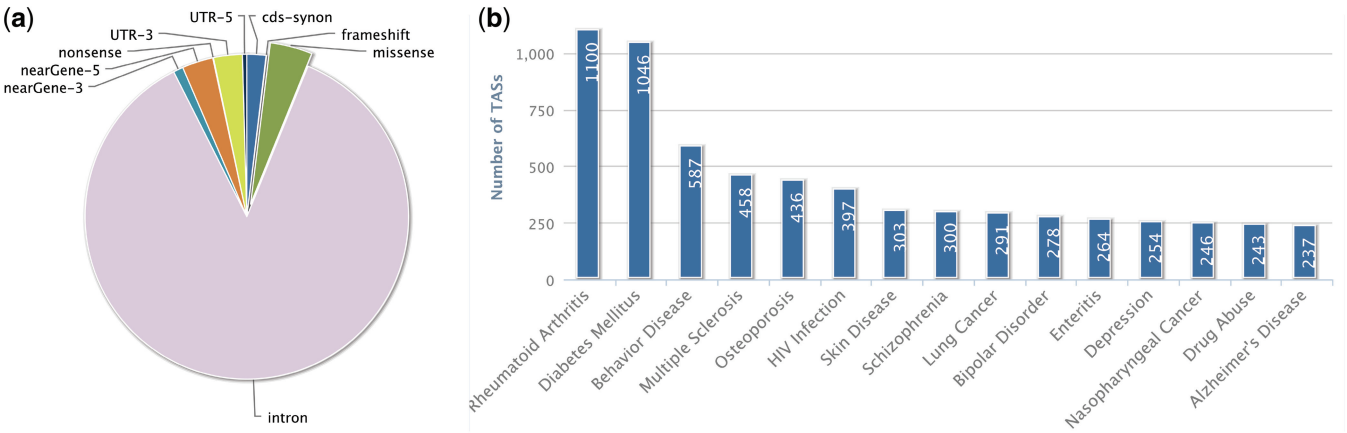

Figure 2. Classifications of GVs from the genic regions and according to the traits/diseases in GWASdb. (**a**) The proportion of GV/gene transcripts with different functional properties in the genic regions (total representing 43.5% of all GVs in GWASdb). (**b**) The Top 15 traits/diseases which have the most significant GVs in database based on DOLite catalog.

the underlying mechanism of SNP–disease association, particularly in the context of pathways and networks. Our database allows users to perform meta-analysis on multiple studies targeting the same disease, defined by a unique term in DOLite or HPO. We used the KGG package (28) to search for enriched pathways or protein–protein interaction networks (PPI). We omitted the disease terms that contained less than 400 GVs because pathway and PPI enrichment analysis need a large dataset of genes.

WEB INTERFACE AND DATA QUERYING

The GWASdb web site provides six straightforward components: Guidance, GWAS overview, Gviewer, DOLite Viewer, HPO Tree Viewer and Customized Page. These help researchers locate and explore the GVs of interest and its related functional annotations.

The guidance page

The GWAS guidance page is the front page of the database. The user should first read this page to get a general idea on the contents and how to use various functions of the database. On the left-upper corner of the page, there is a sliding menu with menu items that the users can start with. If the users want to get all the GVs in the whole genome level, they can click on the 'Overview' item. If they are interested in a particular disease, they can start from either 'DOLITE' or 'HPO' items. If they want to analyze a list of SNPs of they own, they can start from the 'CUSTOMIZED' item.

The GWAS overview page

The GWAS overview page displays a circular GWAS plot showing the global view of the top GVs in each human chromosome. The dots in the plot represent the top two GVs from each study and different colors represent different diseases (Figure 3a). Other information is shown as spectral plots in the inner circles of the plot, such as CNV hotpots, dbSNP density, HapMap density, 1000 genome density and OMIM gene distribution (Figure 3b). By clicking on the ideogram of each chromosome, the user will be presented with a new circular plot displaying a single chromosome showing the top five GVs from each disease. By clicking on a single dot, users will be brought to the Gviewer page and general information of GV will be displayed, such as dbSNP id, *P* value, study source and DOLite catalog number.

The Gviewer

The Gviewer is a web-based genome browser that dynamically displays the different tracks related to the queried GV. Gviewer currently provides four tracks (GV, RefGene, OMIM Gene and DGV) that show the elements around the target GV. More tracks will be added in the future. Users can either click on the arrow buttons or drag the tracks to show the surrounding regions. By clicking elements on the track, users can get detailed information in a popup message box. When a GV is clicked, comprehensive functional annotations of this GV will be displayed on the right pane, which will update with the user actions in the Gviewer. To improve the user experience, selecting different tabs does not switch to another page and waiting time is kept to a minimum because the page loading is asynchronous and the page rendering is progressive.

For example, if a user inputs the dbSNP id (rs437179) in top search bar or by clicking this GV in GWAS overview plot, the user will be automatically forwarded to Gviewer. This shows the GV location in the gene body of SKIV2L, an OMIM gene (600478), together with copy number variants. By clicking on each annotation tab in the right pane, users will obtain the following detailed functional annotations about this GV: (i) it is associated with rheumatoid arthritis (*P* value of 6.15E-20); (ii) it was reported in HapMap and 1000 genome project with average heterozygosity of 0.39; (iii) it has an miRNA (hsa-mir-1236) located in its upstream region; (iv) its two alleles significantly change the transcriptional factor binding site affinities (transcriptional factors: LM105 and GAMYB); (v) it is a non-synonymous SNP; (vi) it is located in the conserved region undergoing positive selection; (vii) it is associated with the differential

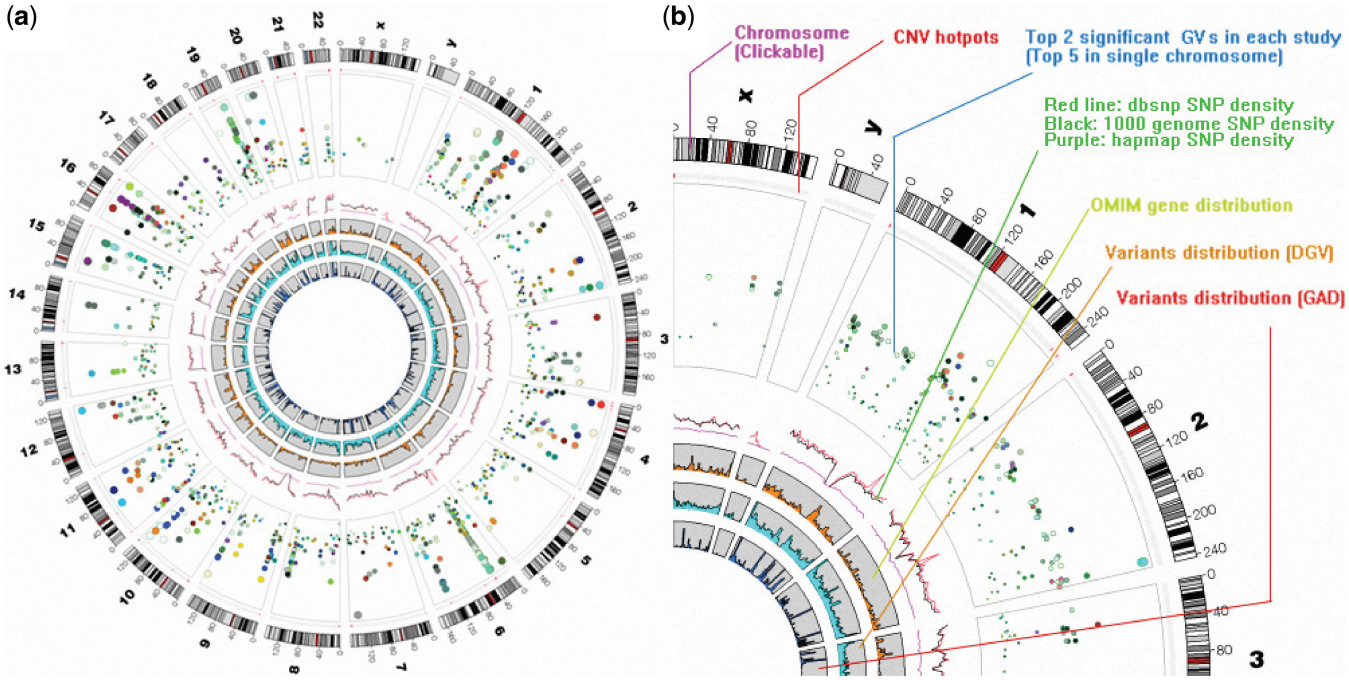

Figure 3. Illustration of the circular GWAS plot. (**a**) Overview of the circular GWAS plot, dots show the top two GVs for each study. (**b**) A description of each of the components in the plot.

co-expression between two genes (DEFB4 and OAS1); (viii) it has extensive variants and diseases association (OMIM: 600478; DGV: 3602, 36507; GAD: 557471, 557472, 557473) (see Supplementary Figure S1).

The DOLite viewer and HPO viewer

To demonstrate the disease/trait classifications of these GVs, we provide the DOLite viewer and HPO viewer. GWASdb displays an interactive Manhattan Plot viewer for easy visualization of GVs mapped to a DOLite node or HPO tree. By selecting each disease or phenotype node, a Manhattan Plot will be instantly drawn on the left pane, with each dot representing a GV. The detailed information on all GVs associated with this disease is simultaneously shown on the right pane. Users can hover the mouse over the GV dot to view a brief description of the GV. When the GV dot is clicked, detailed information will be highlighted on the right pane. By clicking the arrow icon on the highlighted information, the user can continue to the Gviewer page to see the detailed functional annotation of this GV. We also provide pathway and PPI analysis for DOLite terms or HPO nodes that total more than 400 GVs. Two additional tabs can be accessed for gene-based pathway analysis calculated from the KGG package (28) and PPI network analysis rendered by Cytoscape (29) (Supplementary Figure S2).

Searching the GWASdb database

On the front page, users can perform a quick search in any of the five search categories of dbSNP id, gene symbol, chromosome region, DOLite and HPO phenotype terms. The system will show instant hints messages when the user only inputs part search terms or show alert messages if the search term is not recognized. After clicking the 'Go' button, the server will display different views depending on which search category was selected. For example, if dbSNP id was queried, the system will display the highlighted SNP in the Gviewer pane together with comprehensive annotations on the right pane. If the SNP id is in an older version format, the system will automatically convert it to the latest version and process the query. For gene or genomic region searches, the system will show all GVs in that region in the Gviewer pane together with literature information on the right pane. For disease or phenotype queries, a Manhattan Plot will be displayed on the left pane. The user can then click on a particular GV on the plot and the system will display the Gviewer page with that GV highlighted.

The customized page

This customized page allows users to study a list of GVs of their choice. The users will input the list of GVs and select their disease of interest, either as a DOLite term or a HPO node. The server will search our local database for all the SNPs associated with this disease and compare them with the input GVs. A hypergeometric test will be performed to test whether the input GVs have any significant overlap with the GVs in the database. The overlapping and non-overlapping GVs will be displayed in different colors in a Manhattan Plot. By clicking on the dots on the plot, users can further explore the functional annotations of each GV.

Database implementation and downloading

GWASdb is a web-based query tool designed with Service-oriented architecture (SOA). We used jQuery and Raphaël JavaScript frameworks as the frontend to build Gviewer, which ensures high usability of web pages, and we used MySql as the backend database. Database sharding is used to handle the large amount of SNP data. To facilitate the communication with other servers, we have provided web service interfaces for machine-based large-scale data retrieval, which were built using Apache CXF technology (see Supplementary Data). All the functional annotations in the Gview page can be downloaded in batch, by clicking the 'Get All Information in JSON' on the right panel of the Gview page.

DISCUSSION

The GWASdb database can satisfy the demands of the scientific community for the exploration of ever increasing amounts of GWAS data. Many published bioinformatics tools have targeted functional annotations of GVs. We performed a function-oriented comparison with existing tools (see Supplementary Table S2). Using rich web application techniques, GWASdb offers great convenience to researchers for analysis of their GWAS data. Researchers can quickly locate and fetch the GVs of interest and examine the genetic information and functional annotations in great detail. Furthermore, they can explore pathway and PPI networks in the context of disease-oriented meta-analysis. This platform combined with other resources will be an effective tool to study the underlying disease mechanism in GVs. The GWASdb integrative database portal will be a valuable resource for researchers and clinicians.

The GWASdb focuses on specific features and functions of GWAS GVs and their disease classifications. The GWASdb database has collected GVs from six resources so far (NHGRI GWAS Catalog, Johnson and O'Donnell, dbGaP PheGenI, GAD, GWASCentral and PharmGKB). Due to inconsistency in data formats and difficulty of data curation, we did not delve into the experimental and sample description of each GWAS, such as population-related information, individual ratio, geographic region and mode of recruitment. Instead, we provide PubMed links for each GV in our database so that users can easily trace the information from the original publications. Since our purpose was to integrate potentially useful GVs from the literature, we used a predefined cutoff ($P < 1.0 \times 10^{-3}$) as our curation threshold. This cutoff was used because we found most reported moderate SNPs have GWAS significance between 10^{-2} and 10^{-4} (19,30). Nevertheless, lowering the P value cutoff will inevitably increase our false positives. The users can use the 'customized' page to hand pick the GVs of interest. There are experimental methods that can reduce the false positives, for example, validation of GWAS results from an independent cohort, or functional study. Computational methods can also be used to reduce the false positives. For example, it was recently reported that trait/disease-associated GVs are more likely to be expression Quantitative Trait Locus (eQTL). We can use eQTLs to filter the false positive and reveal the true association profile of the study (31).

With the advent of personal genome sequencing projects such as the 1000 Genomes Project, many novel mutations and disease-causing loci will be discovered in the near future. We will constantly recruit new GVs into our database as new GWAS data become available. At the same time, we will incorporate new bioinformatics algorithms and tools to improve the accuracy of functional annotations. In the next stage, we will incorporate SNPs that are not found by GWAS studies, but are in close Linkage Disequilibrium (LD) with the SNPs in GWASdb. This will greatly enhance the utility of this database because there are disease-causing GVs that were not covered by GWAS arrays. Besides, we also aim to collect data from important genome regions such as eQTLs, long non-coding RNA and DNA methylation sites in the next version of GWASdb, because SNPs in those regions may pose positive or negative effects on gene regulation. We will add more tracks to the Gviewer page to allow users to view more functional elements, such as SNP density, haplotype plot and important regulators. For GV annotation, we plan to integrate more data sources or pre-compute the functional predictions using recognized algorithms. The GWASdb database is freely available at http://jjwanglab.org/gwasdb and will be updated frequently.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–2, Supplementary Tables 1–2.

ACKNOWLEDGEMENTS

We thank Kevin Mao and Tina Yuen of The Royal College of Surgeons in Ireland for their assistance in data curation.

FUNDING

The Small Project Fund (201007176262) of the University of Hong Kong; Research Grants Council of Hong Kong (781511M, 778609M, N_HKU752/10); Food and Health Bureau of Hong Kong (10091262); The intramural research program of the National Cancer Institute (NCI), NIH, USA. Funding for open access charge: Research Grants Council (781511M) of Hong Kong.

Conflict of interest statement. None declared.

REFERENCES

1. The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
2. Altshuler,D.M., Gibbs,R.A., Peltonen,L., Dermitzakis,E., Schaffner,S.F., Yu,F., Bonnen,P.E., de Bakker,P.I., Deloukas,P., Gabriel,S.B. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.

3. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
4. Johnson,A.D. and O'Donnell,C.J. (2009) An open access database of genome-wide association results. *BMC Med. Genet.*, **10**, 6.
5. Thorisson,G.A., Lancaster,O., Free,R.C., Hastings,R.K., Sarmah,P., Dash,D., Brahmachari,S.K. and Brookes,A.J. (2009) HGVbaseG2P: a central genetic association database. *Nucleic Acids Res.*, **37**, D797–D802.
6. Mailman,M.D., Feolo,M., Jin,Y., Kimura,M., Tryka,K., Bagoutdinov,R., Hao,L., Kiang,A., Paschall,J., Phan,L. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, **39**, 1181–1186.
7. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.
8. Yu,W., Gwinn,M., Clyne,M., Yesupriya,A. and Khoury,M.J. (2008) A navigator for human genome epidemiology. *Nat. Genet.*, **40**, 124–125.
9. Paananen,J., Ciszek,R. and Wong,G. (2010) Varietas: a functional variation database portal. *Database*, **2010**, baq016.
10. Xu,H., Gregory,S.G., Hauser,E.R., Stenger,J.E., Pericak-Vance,M.A., Vance,J.M., Zuchner,S. and Hauser,M.A. (2005) SNPselector: a web tool for selecting SNPs for genetic association studies. *Bioinformatics*, **21**, 4181–4186.
11. Wang,P.L., Dai,M.H., Xuan,W.J., McEachin,R.C., Jackson,A.U., Scott,L.J., Athey,B., Watson,S.J. and Meng,F. (2006) SNP Function Portal: a web database for exploring the function implication of SNP alleles. *Bioinformatics*, **22**, E523–E529.
12. Lee,P.H. and Shatkay,H. (2008) F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Res.*, **36**, D820–D824.
13. Shen,T.H., Carlson,C.S. and Tarczy-Hornoch,P. (2009) SNPit: a federated data integration system for the purpose of functional SNP annotation. *Comput. Methods Programs Biomed.*, **95**, 181–189.
14. Pico,A.R., Smirnov,I.V., Chang,J.S., Yeh,R.F., Wiemels,J.L., Wiencke,J.K., Tihan,T., Conklin,B.R. and Wrensch,M. (2009) SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system. *Nucleic Acids Res.*, **37**, D803–D809.
15. Chelala,C., Khan,A. and Lemoine,N.R. (2009) SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, **25**, 655–661.
16. Gamazon,E.R., Zhang,W., Konkashbaev,A., Duan,S.W., Kistner,E.O., Nicolae,D.L., Dolan,M.E. and Cox,N.J. (2010) SCAN: SNP and copy number annotation. *Bioinformatics*, **26**, 259–262.
17. Fong,C., Ko,D.C., Wasnick,M., Radey,M., Miller,S.I. and Brittnacher,M. (2010) GWAS analyzer: integrating genotype, phenotype and public annotation data for genome-wide association study analysis. *Bioinformatics*, **26**, 560–564.
18. Wang,J.B., Ronaghi,M., Chong,S.S. and Lee,C.G.L. (2011) pfSNP: an integrated potentially functional SNP resource that facilitates hypotheses generation through knowledge syntheses. *Hum. Mutat.*, **32**, 19–24.
19. Qin,H.D., Shugart,Y.Y., Bei,J.X., Pan,Q.H., Chen,L., Feng,Q.S., Chen,L.Z., Huang,W., Liu,J.J., Jorgensen,T.J. *et al.* (2011) Comprehensive pathway-based association study of DNA repair gene variants and the risk of nasopharyngeal carcinoma. *Cancer Res.*, **71**, 3000–3008.
20. Altman,R.B. (2007) PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat. Genet.*, **39**, 426.
21. Wang,J.W., Zhang,S.L., Schultz,R.M. and Tseng,H. (2006) Search for basonuclin target genes. *Biochem. Biophys. Res. Commun.*, **348**, 1261–1271.
22. Qin,J., Li,M.J., Wang,P., Zhang,M.Q. and Wang,J. (2011) ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. *Nucleic Acids Res.*, **39**, W430–W436.
23. Zhang,G., Chen,X., Chan,L., Zhang,M., Zhu,B., Wang,L., Zhu,X., Zhang,J., Zhou,B. and Wang,J. (2011) An SNP selection strategy identified IL-22 associating with susceptibility to tuberculosis in Chinese. *Sci. Rep.*, **1**, 20.
24. Li,M.J., Sham,P.C. and Wang,J.W. (2010) FastPval: a fast and memory efficient program to calculate very low P-values from empirical distribution. *Bioinformatics*, **26**, 2897–2899.
25. Wang,J.W. and Hannenhalli,S. (2006) A mammalian promoter model links cis elements to genetic networks. *Biochem. Biophys. Res. Commun.*, **347**, 166–177.
26. Du,P., Feng,G., Flatow,J., Song,J., Holko,M., Kibbe,W.A. and Lin,S.M. (2009) From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations. *Bioinformatics*, **25**, i63–i68.
27. Robinson,P.N., Kohler,S., Bauer,S., Seelow,D., Horn,D. and Mundlos,S. (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, **83**, 610–615.
28. Li,M.X., Sham,P.C., Cherny,S.S. and Song,Y.Q. (2010) A knowledge-based weighting framework to boost the power of genome-wide association studies. *PLoS One*, **5**, e14480.
29. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.
30. Adeyemo,A., Gerry,N., Chen,G., Herbert,A., Doumatey,A., Huang,H., Zhou,J., Lashley,K., Chen,Y., Christman,M. *et al.* (2009) A genome-wide association study of hypertension and blood pressure in African Americans. *PLoS Genet.*, **5**, e1000564.
31. Nicolae,D.L., Gamazon,E., Zhang,W., Duan,S., Dolan,M.E. and Cox,N.J. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.*, **6**, e1000888.
32. Saccone,S.F., Quan,J., Mehta,G., Bolze,R., Thomas,P., Deelman,E., Tischfield,J.A. and Rice,J.P. (2011) New tools and methods for direct programmatic access to the dbSNP relational database. *Nucleic Acids Res.*, **39**, D901–D907.
33. Pennacchio,L.A., Visel,A., Minovitsky,S. and Dubchak,I. (2007) VISTA Enhancer Browser—A database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.
34. Cui,Y., Bao,L. and Zhou,M. (2008) CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators. *Nucleic Acids Res.*, **36**, D83–D87.
35. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
36. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
37. Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, D77–D82.
38. Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
39. Betel,D., Koppal,A., Agius,P., Sander,C. and Leslie,C. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.*, **11**, R90.
40. Yang,J.O., Kim,W.Y. and Bhak,J. (2009) ssSNPTarget: genome-wide splice-site single nucleotide polymorphism database. *Hum. Mutat.*, **30**, E1010–E1020.
41. Liu,X., Jian,X. and Boerwinkle,E. (2011) dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.*, **32**, 894–899.
42. Cheng,F., Chen,W., Richards,E., Deng,L. and Zeng,C. (2009) SNP@Evolution: a hierarchical database of positive selection on the human genome. *BMC Evol. Biol.*, **9**, 221.
43. Wang,Y., Joseph,S.J., Liu,X., Kelly,M. and Rekaya,R. (2011) SNPxGE2: a database for human 3-way SNP-expression associations. *Nature Precedings* (doi:10.1038/npre.2011.5704.1; epub ahead of print).