GWAS Atlas: a curated resource of genome-wide variant-trait associations in plants and animals

Dongmei Tian¹,²,†, Pei Wang¹,²,³,†, Bixia Tang¹,²,†, Xufei Teng¹,²,³, Cuiping Li¹,², Xiaonan Liu¹,²,⁴, Dong Zou¹,², Shuhui Song¹,²,³,⁵,* and Zhang Zhang¹,²,³,⁴,⁵,*

¹National Genomics Data Center, Beijing 100101, China, ²BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, ³College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China, ⁴School of Future Technology, University of Chinese Academy of Sciences, Beijing 100049, China and ⁵CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

Received August 15, 2019; Revised September 06, 2019; Editorial Decision September 13, 2019; Accepted September 22, 2019

ABSTRACT

GWAS Atlas (https://bigd.big.ac.cn/gwas/) is a manually curated resource of genome-wide variant-trait associations for a wide range of species. Unlike existing related resources, it features comprehensive integration of a high-quality collection of 75 467 variant-trait associations for 614 traits across 7 cultivated plants (cotton, Japanese apricot, maize, rape, seed, rice, sorghum and soybean) and two domesticated animals (goat and pig), which were manually curated from 254 publications. We integrated these associations into GWAS Atlas and presented them in terms of variants, genes, traits, studies and publications. More importantly, all associations and traits were annotated and organized based on a suite of ontologies (Plant Trait Ontology, Animal Trait Ontology for Livestock, etc.). Taken together, GWAS Atlas integrates high-quality curated GWAS associations for animals and plants and provides user-friendly web interfaces for data browsing and downloading, accordingly serving as a valuable resource for genetic research of important traits and breeding application.

INTRODUCTION

Genome-wide association study (GWAS) is a key technique for exploiting the genetic basis of complex traits and diseases by detecting genotype-phenotype associations in a group of individuals or natural inbred lines (1). As a result, GWAS has been widely applied in model organisms, primarily in human (2,3) and Arabidopsis (4). Nowadays, with the rapid development of phenotyping and genotyping technologies, a number of high-quality genotype-phenotype associations have also been identified in plants and animals, including maize, rice, sorghum, cotton, soybean, goat, pig, etc. (5,6). Taking corn as an example, diverse traits ranging from molecular to cellular (i.e. metabolites) and from individual morphological scale (agronomic, yield or reproductive characteristics) to the interaction with different environmental factors (biotic or abiotic stress tolerance), have been comprehensively studied and associated with different genetic variants (5,7). Therefore, a resource that integrates GWAS associations, as well as their associated information, is of fundamental significance to systematically capture the whole picture of genotype-phenotype associations and improve our understanding of genetic architecture of complex quantitative traits.

Over the past decade, several resources, such as GWASdb (8,9), GWAS Catalog (3,10,11), GWASCentral (12) and AraGWAS Catalog (13), have been developed to provide publicly available GWAS associations. Among them, GWAS Catalog and AraGWAS Catalog are two popular representatives. GWAS Catalog, a dedicated resource for human that includes a total of 149 855 significant associations ($P < 1 \times 10^{-5}$) derived from 7230 studies in 4085 publications (as of August 2019) (3), has been widely used by the global community for deciphering genetic basis and molecular mechanisms of human phenotypes and diseases. AraGWAS Catalog (13), a curated and standardized GWAS associations database for Arabidopsis, has integrated more than 167 traits and 222 000 variant-trait associations, making it an important and useful resource for the Arabidopsis community. Despite this, existing resources focused mainly on human and other model organisms and did not comprehensively integrate GWAS associations identified in both crops and domesticated animals. Therefore, a specialized resource that houses comprehensive GWAS associations as well as...
their associated information in a variety of plants and animals is highly desirable.

Here, we present GWAS Atlas (https://bigd.big.ac.cn/gwas), a curated resource incorporating high-quality GWAS associations with particular focus on plants and animals. In the current release, it houses a total of 75 467 GWAS associations in seven cultivated plants and two domesticated animals manually curated from 254 publications. Moreover, all associations and traits are annotated and organized based on a suite of ontologies. To facilitate data access and query, GWAS Atlas is equipped with friendly web interfaces to provide genetic markers and genes that have significant associations ($P < 10^{-3}$) with specific traits of interest. Thus, GWAS Atlas would function as a valuable resource for studying complex agronomic traits and conducting genomic breeding applications in plants and animals.

**IMPLEMENTATION**

GWAS Atlas is implemented using MySQL (http://www.mysql.org; a free and popular relational database management system) and Apache Tomcat Server (http://tomcat.apache.org; an open source software implementation of Java Servlet and Java Server Pages). Web user interfaces are developed using JSP (Java Server Pages; a technology facilitating rapid development of dynamic web pages based on the Java programming language), HTML5, CSS3, AJAX (Asynchronous JavaScript and XML; a set of web development techniques to create asynchronous applications without interfering with the display and behavior of the existing page), JQuery (a cross-platform and feature-rich JavaScript library; http://jquery.com, version 3.2.1) as well as Bootstrap (an open source toolkit for developing web projects with HTML, CSS and JS; https://getbootstrap.com, version 3.3.7). For dynamic data visualization, ECharts (a declarative framework for rapid construction of web-based visualization; http://echarts.baidu.com, version 4.1.0) is adopted to generate interactive charts.

**DATA CURATION**

To provide high-quality information curated from GWAS publications, we set up a standardized curation process involving four major steps, viz., literature search, information retrieval, integration & annotation and database construction (Figure 1). Briefly speaking, first, we perform literature search in PubMed using species name and GWAS as keywords and accordingly obtain a total of 1580 publications. Among them, 1767 publications published after 2009 are retained. Publications are eligible for inclusion in GWAS Atlas if they contain significant GWAS associations with necessary description on biological traits. Consequently, a total of 254 publications are qualified and their basic bibliographic information (e.g. title, journal, year, citation) are automatically obtained through Europe PMC API (https://europepmc.org/developers/) (14,15). Then, we manually curate the study and genotype-to-phenotype (G2P) association information from publications. As one publication may contain multiple studies with different experimental designs, we record species name, sampling spot, year, condition, population, sample size, genotyping technology, association model, association number, and PMID for each study. Regarding GWAS association, we collect species name, genome version, genomic position, variant ID, traits, GWAS association $P$-value, $R^2$ and mapped genes. Considering the possible inconsistency of reference genome versions used in different studies, all genomic variants for a given species are annotated based on the latest version. Finally, to unify the representation of biological traits, trait entities are mapped to a suite of reference ontologies (PTO, Plant Trait Ontology; ATOL, Animal Trait Ontology for Livestock) (16–18) and species-specific ontology (CO, Crop Ontology) (19) using the ‘term search’ in Plantome API and Livestock Ontologies. Since not all curated traits are included in existing ontologies, we additionally establish PPTO (Plant Phenotype and Trait Ontology) and APTO (Animal Phenotype and Trait Ontology) by integrating more comprehensive terms based on Open Biological and Biomedical Ontologies (OBO) format.

**DATABASE CONTENT AND USAGE**

Based on the standardized curation process, GWAS Atlas integrates a high-quality collection of 75 467 variant-trait associations for 614 traits across seven cultivated plants and two domesticated animals, which were manually curated from 254 publications (Table 1). To facilitate users in browsing these data, GWAS Atlas provides eight modules, where data are organized and presented in terms of species, associations, variants, genes, traits, ontologies, studies and publications, respectively.

The ‘Associations’ module provides a comprehensive overview of associated hits ($P < 10^{-3}$) for each species, where data are organized in a tabular table (Figure 2A) and displayed in a heatmap at a chromosome-scale (Figure 2B). Each association contains its variant ID, chromosomal position, consequence type(s), mapped gene(s), associated trait, $P$-value and PMID. And links to external web sites are provided to help users easily access more details about variants and genes. Besides, a set of filters in light of $P$-value, trait and variant position are provided to help users narrow down the browsing list. Moreover, associations could also be sorted by a variety of keywords, including VarID, traits, species, $P$-Value and $R^2$%. In the ‘Variants’ (Figure 2C) and ‘Genes’ modules (Figure 2D), we summarize all associated hits detected in non-redundant variants and genes (or in close proximity to genes) and group the results by variant ID and gene name, respectively. Detailed information can be obtained by clicking on any variant ID or gene ID. Through sorting by trait or association count, users can swiftly get shared genes or sites with multiple different effects.

The ‘Traits’ module provides an overview of all collected traits in GWAS Atlas. For each trait, both general details (e.g. trait label, trait ID, description) and summary information (e.g. association, study, publication) are recorded in the trait table (Figure 3A). As one trait may have multiple associated variants and genes, users could easily access these data by clicking on their hyperlinks. Besides, these traits are sortable by trait label, trait ID and numbers of associations, studies and publications. According to the current collec-
Figure 1. Data curation process adopted by GWAS Atlas.

Table 1. Data statistics in GWAS Atlas (as of 10 August 2019)

| Species                  | # Publications | # Studies | # Traits | # Associations | # Variants | # Genes |
|--------------------------|----------------|-----------|----------|----------------|------------|---------|
| Cotton (*Gossypium hirsutum*) | 4              | 23        | 24       | 21,955         | 6,115      | 991     |
| Japanese apricot (*Prunus mume*) | 1              | 2         | 9        | 1,740          | 14,32      | 541     |
| Maize (*Zea mays*)        | 86             | 308       | 205      | 28,310         | 22,969     | 13,654  |
| Rape seed (*Brassica napus*) | 26             | 89        | 67       | 2,396          | 1,709      | 1,791   |
| Rice (*Oryza sativa*)     | 96             | 456       | 284      | 19,524         | 14,528     | 12,803  |
| Sorghum (*Sorghum bicolor*) | 13             | 52        | 41       | 754            | 652        | 688     |
| Soybean (*Glycine max*)   | 15             | 46        | 40       | 422            | 337        | 354     |
| Goat (*Capra hircus*)     | 4              | 6         | 12       | 40             | 40         | 9       |
| Pig (*Sus scrofa*)        | 9              | 10        | 35       | 326            | 284        | 122     |

Discussion and Future Directions

GWAS Atlas incorporates a large number of high-quality variant-trait associations in multiple plants and animals through manual curation. It equips with friendly web interfaces for browse, search and visualization, thus enabling users to easily maneuver the GWAS associations and uncover molecular mechanisms underlying complex traits. Therefore, GWAS Atlas would be helpful for fully captur-
Figure 2. Screenshots for Associations (A), GWAS heatmap across chromosomes (B), Variants (C) and Genes (D). Variant-trait associations in cotton are used to depict the GWAS heatmap, where each row represents a trait and each dot illustrates an associated hit. The $\log(P)$ is adopted to reflect the statistical significance.
ing pleiotropic loci and better understanding the similarities and differences in genetic mechanisms between different species and/or traits. So far, the current release of GWAS Atlas includes 9 species, 614 traits and 75,467 variant-trait associations. Accordingly, future directions are integration of more GWAS findings from a broader range of species and continuously updating associations as well as related information. To further evaluate the effect of variants on other biological processes and facilitate the discovery of potential molecular mechanism for these variant sites, we plan to add additional curated information (such as RNA binding protein and microRNA binding site). Besides, as several ontologies used in this study have not been determined their parent-child relationships, we will further optimize and improve their relationships according to the ontology standards and by reference of PTO and ATO. Moreover, we plan to develop the submission functionality to allow users to submit their own data. Meanwhile, we call for worldwide

Figure 3. Screenshots for Traits (A), Ontologies (B), Studies (C) and Publications (D).
collaborations to work together to build GWAS Atlas into a valuable resource covering more comprehensive associations and traits across a wider range of plants and animals.

ACKNOWLEDGEMENTS

We thank members of the National Genomics Data Center for reporting bugs and sending comments.

FUNDING

Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) [XDA19050302, XDB13040500]; National Key Research & Development Program of China [2017YFC0907502]; National Natural Science Foundation of China [31871328, 31771465, 31671360, 31571358]; International Partnership Program of the CAS [153F11KYSB20160008]; 15th Third-year Informatization Plan of the CAS [XXH13505-05 to Z.Z.]; Key Technology Talent Program of the CAS; The 100 Talent Program of the Chinese Academy of Sciences; K.C. Wong Education Foundation [to Z.Z.]; The Youth Innovation Promotion Association of Chinese Academy of Sciences [2017141 to S.S.]. Funding for open access charge: The Promotion Association of Chinese Academy of Sciences Education Foundation [to Z.Z.]; The Youth Innovation Program of the Chinese Academy of Sciences; K.C. Wong Technology Talent Program of the CAS; The 100 Talent Informatization Plan of CAS [XDA19050302, XDB13040500]; National Key Research & Development Program of China [XDA19050302]; Key Technology Talent Program of the CAS; The 100 Talent Program of the Chinese Academy of Sciences; K.C. Wong Education Foundation [to Z.Z.]; The Youth Innovation Promotion Association of Chinese Academy of Sciences. Funding for reporting bugs and sending comments. We thank members of the National Genomics Data Center for reporting bugs and sending comments.

REFERENCES

1. Bush,W.S. and Moore,J.H. (2012) Chapter 11: Genome-wide association studies. PLoS Comput. Biol., 8, e1002822.
2. Klein,R.J., Zeiss,C., Chew,E.Y., Tsai,J.Y., Sackler,R.S., Haynes,C., Henning,A.K., SanGiovanni,J.P., Mane,S.M., Mayne,S.T. et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science, 308, 385–389.
3. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Henning,A.K., SanGiovanni,J.P., Mane,S.M., Mayne,S.T. et al. (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res., 47, D1005–D1012.
4. 1001 Genomes Consortium. (2016) 1,135 Genomes reveal the global pattern of polymorphism in arabidopsis thaliana. Cell, 166, 481–491.
5. Liu,H.J. and Yan,J. (2019) Crop genome-wide association study: a harvest of biological relevance. Plant J., 97, 8–18.
6. Song,S.H., Tian,D.M., Zhang,Z., Hu,S.N. and Yu,J. (2019) Rice genomics: over the past two decades and into the future. Genom. Proteom. Bioinf., 16, 397–404.
7. Xiao,Y., Liu,H., Wu,L., Warburton,M. and Yan,J. (2017) Genome-wide association studies in maize: Praise and stargaze. Mol. Plant., 10, 359–374.
8. Li,M.J., Liu,Z., Wang,P., Wong,M.P., Nelson,M.R., Kocher,J.P., Yeager,M., Sham,P.C., Chanock,S.J., Xia,Z. et al. (2016) GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res., 44, D869–D876.
9. Li,M.J., Wang,P., Liu,X., Lim,E.L., Wang,Z., Yeager,M., Wong,M.P., Sham,P.C., Chanock,S.J. and Wang,J. (2012) GWASdb: a database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res., 40, D1047–D1054.
10. MacArthur,J., Bowler,E., Cerezo,M., Gil,L., Hastings,E., Junkins,H., McMahon,A., Morales,J. et al. (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res., 45, D896–D901.
11. Welte,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,J. et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res., 42, D1001–D1006.
12. Beck,T., Hastings,R.K., Gollapudi,S., Free,R.C. and Brookes,A.J. (2014) GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. Eur. J. Hum. Genet., 22, 949–952.
13. Togninalli,M., Seren,U., Meng,D., Fitz,J., Nordborg,M., Weigel,D., Tannenbaum,S., Korte,A. and Grimm,D.G. (2016) The AraGWAS Catalog: a curated and standardized Arabidopsis thaliana GWAS catalog. Nucleic Acids Res., 46, D1150–D1156.
14. Levenchenko,M., Gou,Y., Graef,F., Hamelers,A., Huang,Z., Ide-Smith,M., Iyer,A., Kiliyan,O., Katuri,J., Kim,J.H. et al. (2018) Europe PMC in 2017. Nucleic Acids Res., 46, D1254–D1260.
15. Europe,F.M.C.C. (2015) Europe PMC: a full-text literature database for the life sciences and platform for innovation. Nucleic Acids Res., 43, D1042–D1048.
16. Cooper,L., Meier,A., Laporte,M.A., Elser,J.L., Mungall,C., Sinn,B.T., Cavaliere,D., Carbon,S., Dunn,N.A., Smith,B. et al. (2018) The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Res., 46, D1168–D1180.
17. Hulsegge,B., Smits,M.A., te Pas,M.F. and Woelders,H. (2012) Contributions to an animal trait ontology. J. Anim. Sci., 90, 2061–2066.
18. Hughes,L.M., Bao,J., Hu,Z.L., Honavar,V. and Reecy,J.M. (2008) Animal trait ontology: the importance and usefulness of a unified trait vocabulary for animal species. J. Anim. Sci., 86, 1485–1491.
19. Shrestha,R., Matteis,L., Skofic,M., Portugal,A., McLaren,G., Hyman,G. and Arnaud,E. (2012) Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the crop ontology developed by the crop communities of practice. Front Physiol., 3, 326.
20. Song,S., Tian,D., Li,C., Tang,B., Dong,L., Xiao,J., Bao,Y., Zhao,W., He,H. and Zhang,Z. (2018) Genome variation map: a data repository of genome variations in BIG Data Center. Nucleic Acids Res., 46, D944–D949.
21. BIG Data Center Members. (2019) Database resources of the BIG data center in 2019. Nucleic Acids Res., 47, D8–D14.