D868–D872 *Nucleic Acids Research, 2015, Vol. 43, Database issue* *Published online 31 October 2014*
doi: 10.1093/nar/gku1037

VaDE: a manually curated database of reproducible associations between various traits and human genomic polymorphisms

Yoko Nagai[1,†], Yasuko Takahashi[1,†] and Tadashi Imanishi[1,2,*]

[1]Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Kanagawa 259-1193, Japan and [2]Data Management and Integration Team, Molecular Profiling Research Center for Drug Discovery, National Institute of Advanced Industrial Science and Technology, Koto-ku, Tokyo 135-0064, Japan

Received August 15, 2014; Revised October 09, 2014; Accepted October 10, 2014

ABSTRACT

Genome-wide association studies (GWASs) have identified numerous single nucleotide polymorphisms (SNPs) associated with the development of common diseases. However, it is clear that genetic risk factors of common diseases are heterogeneous among human populations. Therefore, we developed a database of genomic polymorphisms that are reproducibly associated with disease susceptibilities, drug responses and other traits for each human population: 'VarySysDB Disease Edition' (VaDE; http://bmi-tokai.jp/VaDE/). SNP-trait association data were obtained from the National Human Genome Research Institute GWAS (NHGRI GWAS) catalog and RAvariome, and we added detailed information of sample populations by curating original papers. In addition, we collected and curated original papers, and registered the detailed information of SNP-trait associations in VaDE. Then, we evaluated reproducibility of associations in each population by counting the number of significantly associated studies. VaDE provides literature-based SNP-trait association data and functional genomic region annotation for SNP functional research. SNP functional annotation data included experimental data of the ENCODE project, H-InvDB transcripts and the 1000 Genome Project. A user-friendly web interface was developed to assist quick search, easy download and fast swapping among viewers. We believe that our database will contribute to the future establishment of personalized medicine and increase our understanding of genetic factors underlying diseases.

INTRODUCTION

Genome-wide association studies (GWASs) have identified numerous single nucleotide polymorphisms (SNPs) that are associated with development of multifactorial diseases, such as coronary artery disease, rheumatoid arthritis, type 2 diabetes mellitus and cancers (1). However, because GWASs use statistical evaluation, we cannot completely eliminate false positives that may contaminate the data. On the other hand, it is becoming clear that genetic risk factors of common diseases are not totally universal, but are heterogeneous among human populations. For example, disease-associated SNPs have different effects and frequencies between different populations, such as European and East Asian populations (2). According to our population-based Rheumatoid Arthritis association database, RAvariome, 30 of 79 rheumatoid arthritis-associated SNPs are unique to East Asian (3).

Existing SNP-trait association databases, such as a catalog of published GWASs of the National Human Genome Research Institute (NHGRI GWAS catalog) (4), GWASdb (5), GWAS Central (6), HuGE Navigator (7), dbGaP (8) and PharmGKB (9), were developed by collecting data from published articles or are repository databases that set up an infrastructure for researchers. However, these databases do not have sufficient information on the human subjects, especially the ancestry of populations examined.

In this study, we introduce the VarySysDB Disease Edition (VaDE) database, which provides various trait-related, human genetic risk information based on an assessment of reproducibility of the association for each human population. The SNP-trait association data were collected from the NHGRI GWAS catalog, RAvariome and from manual curation of the literature. In addition to subject ethnic information from NHGRI GWAS catalog, we curated ancestral population and nationality of subjects from the original articles. Furthermore, functional information of genomic regions for each SNP was integrated to permit the identifica-

*To whom correspondence should be addressed. Tel: +81 463 93 1121 (Ext 2140); Fax: +81 463 93 5418; Email: imanishi@tokai-u.jp
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

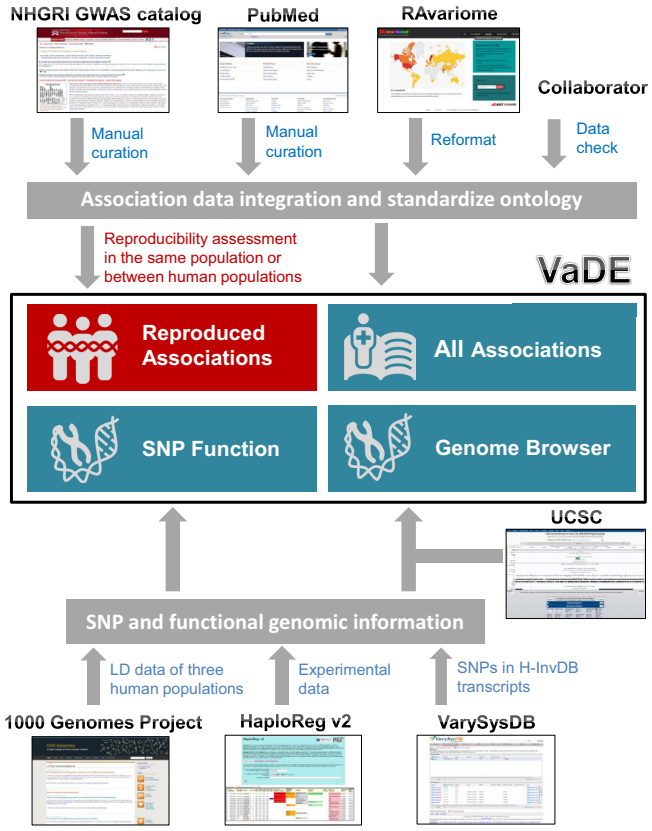

Figure 1. Data flow in VaDE database. The data resources are shown at the top and bottom. VaDE database contents are shown in the middle.

tion of functional SNPs. Experimental data, such as ChIP-seq, DNase I hypersensitivity experiments, regulatory motifs, chromatin state segmentation, RefSeq genes, H-InvDB transcripts and linkage disequilibrium (LD) data in three major human populations were integrated and linked to the SNPs. We also installed a genome browser to visualize these data.

THE VaDE DATABASE

Collection of association data

Figure 1 illustrates the data flow in the VaDE database. There are two types of SNP-trait association data in VaDE. One was collected from a NHGRI GWAS catalog up to 20 January 2014. Detailed information of the literature curation process of the NHGRI GWAS catalog is described at http://www.genome.gov/gwastudies/, which we followed basically. In addition to items from the NHGRI GWAS catalog, we added items such as subject ancestral population, subject nationality, number of studies in the article and number of significant studies reported in the article. NHGRI GWAS catalog data were carefully checked and corrected with reference to the original articles. The association results of the articles used a variety of genetic models; therefore, we integrated associations of allelic or additive models and excluded results of dominant and recessive models. As of October 2014 (VaDE version 1.3), 4169 pieces of data from the NHGRI GWAS catalog have been modified and 169 have been deleted from the 15 542 NHGRI GWAS entries.

In addition, comprehensive manual curation of selected diseases is continuously carried out by the VaDE team. The aim of the project is to collect comprehensive association data from non-European population studies, and assess the data quality and reproducibility mechanically to avoid curator bias. Extracted SNP-trait associations were not limited to those with P-values $<1.0 \times 10^{-5}$ because there is a possibility that this would exclude recent large-scale studies. As of October 2014 (VaDE version 1.3), 7530 association pieces of data from rheumatoid arthritis (based on our previous database, RAvariome) and hypertension were included in the VaDE database. Sixty-seven type 2 diabetes-associated data were provided by collaboration with the Tokyo Medical and Dental University (10). These comprise highly curated datasets, and with continued curation, the amount of data will continue to increase. After data of the NHGRI GWAS catalog and VaDE were merged, 90 duplicated data were excluded from the database.

Standardize ontology

The World Factbook of the CIA and Composition of Regions of United Nations Statistics Division were used to standardize the vocabulary of nations and populations. Subject population information was classified into the following 10 populations; European (including European American, European Australian, European New Zealander, West European, North European, South European and East European), African (including African American), East Asian, South-East Asian, South Asian, West Asian, Caribbean/Central American (such as Caribbean Hispanic, Latino, Caribbean, Hispanic), North American (such as Native American, American Indian, Native Alaskan, Pima Indian, Tohono O'odham Indian, Native Alaskan, Native Indian), South American and Oceanian.

Traits and diseases were classified into the following five categories: disease, multiple diseases, medical trait, general trait and pharmacogenomics. Diseases were further classified based on the WHO *International Classification of Diseases, 10th Revision* by using the Unified Medical Language System. SNP location and nearest gene annotation are based on dbSNP137. A reported gene is that reported in the original article. A related gene is that normalized by the expressions of reported genes from many articles, such that the representation is unique to the SNP.

Reproducibility assessment of SNP-trait association

To assess the reproducibility of a SNP-trait association, the SNP-trait association data were excluded when a P value was not reported, where the association is not significant and where the sample came from multiple populations. For each SNP, the number of total studies and the number of significant results in the total studies were counted when they were tested by other independent samples, such as a replication study. For each combination of population, trait, SNP and risk allele, the total number of significant studies was counted between independent articles to confirm the reproducibility in the population. SNP-trait associations whose reproducibility was confirmed by more than

Figure 2. Screen shot of VaDE contents. (**A**) Reproduced Associations page, (**B**) All Associations page, (**C**) SNP Functional Annotations page and (D) Genome browser.

Table 1. Statistics of the population-based reproducibility assessment (VaDE version 1.3)

Population	Number of associations	Number of unique variants	Number of unique traits
European	2446	2196	288
African	67	56	19
East Asian	557	456	102
South-East Asian	10	10	5
South Asian	26	13	3
West Asian	4	2	2
Caribbean/Central American	22	15	9
North American	0	0	0
South American	8	4	1
Oceanian	0	0	0

one independent study within a literature or between literatures were summarized in the Reproduced Associations page.

Finally, VaDE (version 1.3) provided 15 283 NHGRI GWAS catalog data and 7530 association data from manual curation. By an assessment of reproducibility of each combination of disease, population, variant (SNP, haplotype or HLA allele) and risk allele, 3140 reproducible associations were found within or between human populations. Table 1 shows the statistics of the total number of reproducible associations, unique variants and unique traits for each human population. Only 67 associations were reproducible between two different populations (Supplementary Table S1).

Collection of functional genomic data

Experimental data, such as ChIP-seq, DNase I hypersensitivity experiments, regulatory motifs data from ENCODE project (11), chromatin state segmentation from ENCODE/Broad (12,13), chromatin state segmentation from NIH Roadmap Epigenomics Mapping Consortium (14) and RefSeq gene annotation (15) were downloaded from HaploReg v2 (16). Annotation of H-InvDB transcripts to SNPs was based on VarySysDB (17,18). LD between a SNP and 3000 upstream and downstream SNPs based on the 1000 Genome Project (19) were provided by the Center for Statistical Genetics, University of Michigan. Detailed information for the method is available at http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G.2012-02-14.html. To visualize the location and relation between SNPs and functional genomic regions, we developed a genome browser based on the UCSC Genome Browser (20).

Web interface

The VaDE database comprises four pages: the Reproduced Associations page, the All Associations page, the SNP Functional Annotations page and the Genome Browser page (Figure 2).

On the Reproduced Associations page, a user can search association data by trait/disease name, gene name (gene symbol), SNP ID (dbSNP rs number) or HLA allele name and population name (Figure 2A). By filtering the number of populations, a user can search association data whose reproducibility was confirmed in more than one population. In the left section, a list of the number of studies according to human population is displayed for every record of combination of trait/disease, gene, variant and allele. In the right section, information of selected SNP-trait associations from the latest study is shown as a representative result for each population.

In the All Associations page, manually curated SNP-trait association data are provided (Figure 2B). In the search field, a user can search for the country where the samples were collected. In the left section, in addition to trait/disease, gene, SNP-allele and population, *P*-value of SNP-trait association data and OR/beta, the reported article's PubMed ID and number of studies in the article are provided. The number of studies in the article is not counted when the data is not significant, does not report a *P*-value or is a multiple population analysis.

The SNP Functional Annotations page and the Genome Browser page provide SNP-related functional genomic regions and LD information of European, African and East Asian populations (Figure 2C and D). The Genome Browser page provides the location of SNPs, the LD block of the SNPs, genes and functional genomic region. The LD block shown in the Genome Brower is limited to SNPs in the VaDE All Associations page. The LD block is defined by the farthest SNPs that link to the focus SNP in $r^2 > 0.8$, both upstream and downstream. Using the Genome browser, a user can easily determine the location and distance between a focus SNP and functional genomic regions.

All four pages are linked by hyperlinks. In the Reproduced Associations page, hyperlinks in the trait/disease column, gene name column and SNP-allele column take the user to the All Associations page by searching with the clicked query. For example, if user clicked 'rs887829-T' in the SNP-allele column, the result of searching 'rs887829-T' in the All Associations page would be shown (Figure 2A and B). In the All Associations page, users can jump to the SNP Functional Annotations page by clicking the SNP rs number. Figure 2C show the list of SNPs that correlated with rs887829 ($r^2 > 0.8$ in European, African or Asian populations) and the genomic functional region that overlapped with the SNP location. Clicking the location column of rs887829 takes the user to the Genome Browser page (Figure 2D).

Future perspectives

GWASs are currently producing large amounts of data worldwide, and will continue to do so. GWASs report genomic polymorphisms that are associated with various hu-

man phenotypes in numerous scientific papers. Also, it is becoming clear that not only SNPs but also copy number variations and other structural variants are associated with human phenotypes. The VaDE development team will continue to collect these data, which will be released regularly. However, the VaDE development team alone cannot survey all GWAS papers published in all major journals. Thus, in the near future, we plan to develop a data submission system by which GWAS researchers can submit their own results to VaDE, which may facilitate feedback from the user community.

Currently, VaDE viewers provide users with several hyperlinks to major public databases, such as PubMed and dbSNP (21). In the future, we plan to offer further hyperlinks to external databases concerning human genomic polymorphisms using the Hyperlink Management System (22) that is an automated ID mapping system. This will enable VaDE users to obtain more data about genomic polymorphisms dispersed worldwide, and to analyze them in an integrated manner. Through the development of an integrated database of genomic polymorphisms, VaDE, we will continue to provide the research community with reliable association data to support the development of future preventive medicines and basic research on human phenotypes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We express our thanks to Drs Noriko Sato, Nay Chi Htun and Masaaki Muramatsu of the Tokyo Medical and Dental University for providing us with curated data of type 2 diabetes. We also thank Drs Takato Matsui and Junichi Takeda and members of the Support Center for Medical Research and Education, Tokai University for manual curation of the literature and Kensuke Numakura for valuable discussions and disease classification. Finally, we thank Nobuo Obi, Takuya Habara and Kentaro Mamiya for technical support for the database development.

FUNDING

JSPS KAKENHI [258055, 268046]. Funding for openaccess charge: Tokai University School of Medicine.
Conflict of interest statement. None declared.

REFERENCES

1. Visscher,P.M., Brown,M.A., McCarthy,M.I. and Yang,J. (2012) Five years of GWAS discovery. *Am. J. Hum. Genet.*, **90**, 7–24.
2. Okada,Y., Wu,D., Trynka,G., Raj,T., Terao,C., Ikari,K., Kochi,Y., Ohmura,K., Suzuki,A., Yoshida,S. *et al.* (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, **506**, 376–381.
3. Nagai,Y. and Imanishi,T. (2013) RAvariome: a genetic risk variants database for rheumatoid arthritis based on assessment of reproducibility between or within human populations. *Database (Oxford)*, **2013**, bat073.
4. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
5. 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.
6. 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.
7. Yu,W., Gwinn,M., Clyne,M., Yesupriya,A. and Khoury,M.J. (2008) A navigator for human genome epidemiology. *Nat. Genet.*, **40**, 124–125.
8. Tryka,K.A., Hao,L., Sturcke,A., Jin,Y., Wang,Z.Y., Ziyabari,L., Lee,M., Popova,N., Sharopova,N., Kimura,M. *et al.* (2014) NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res.*, **42**, D975–D979.
9. McDonagh,E.M., Whirl-Carrillo,M., Garten,Y., Altman,R.B. and Klein,T.E. (2011) From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomark. Med.*, **5**, 795–806.
10. Sato,N., Htun,N.C., Daimon,M., Tamiya,G., Kato,T., Kubota,I., Ueno,Y., Yamashita,H., Fukao,A., Kayama,T. *et al.* (2014) Likelihood ratio-based integrated personal risk assessment of type 2 diabetes. Endocrinol. J., EJ14-0271.
11. ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
12. Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol.*, **28**, 817–825.
13. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
14. Chadwick,L.H. (2012) The NIH Roadmap Epigenomics Program data resource. *Epigenomics*, **4**, 317–324.
15. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
16. Ward,L.D. and Kellis,M. (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, **40**, D930–D934.
17. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, e162.
18. Shimada,M.K., Matsumoto,R., Hayakawa,Y., Sanbonmatsu,R., Gough,C., Yamaguchi-Kabata,Y., Yamasaki,C., Imanishi,T. and Gojobori,T. (2009) VarySysDB: a human genetic polymorphism database based on all H-InvDB transcripts. *Nucleic Acids Res.*, **37**, D810–D815.
19. Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
20. Karolchik,D., Barber,G.P., Casper,J., Clawson,H., Cline,M.S., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, **42**, D764–D770.
21. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
22. Imanishi,T. and Nakaoka,H. (2009) Hyperlink Management System and ID Converter System: enabling maintenance-free hyperlinks among major biological databases. *Nucleic Acids Res.*, **37**, W17–W22.