Journal of Human Genetics (2009) 54, 543–546

www.nature.com/jhg

SHORT COMMUNICATION

Genome-wide association database developed in the Japanese Integrated Database Project

Asako Koike[1], Nao Nishida[2], Ituro Inoue[3], Shoji Tsuji[4] and Katsushi Tokunaga[2]

The establishment of high-throughput single-nucleotide polymorphism (SNP)-typing technologies has enabled astonishing progress to be made in genome-wide association studies (GWAS), and various novel genetic factors associated with complex diseases have been discovered. Our organization has created a public repository database (DB) to achieve a continuous and intensive management of GWAS data and to facilitate data sharing among researchers. In the GWAS DB, information on study design, quality control protocols, allele frequencies, genotype frequencies and statistical genetic analysis results are stored as publicly available data and can be accessed freely, whereas individual genotyping data and raw data are stored as restricted data and can only be accessed with authorization. All data are presented by a graphic viewer, which is designed to be user friendly for researchers who are not familiar with GWAS to accelerate disease-related studies. Furthermore, the DB allows users to compare various study results obtained by different institutions and on different platforms. The same data are also managed as a distributed annotation system to call up useful data from other DBs and to superimpose them on the GWAS data for help in interpretation. The DB is accessible at https://gwas.lifesciencedb.jp/.
Journal of Human Genetics (2009) **54,** 543–546; doi:10.1038/jhg.2009.68; published online 24 July 2009

Keywords: database; genome-wide association; SNP

INTRODUCTION

The accomplishment of sequencing of the entire human genome[1,2] and the HapMap project,[3] coupled with the development of cost-effective high-throughput dense single-nucleotide polymorphism (SNP)-typing techniques, have enabled a genome-wide exploration of various complex disease-associated variants. Currently, the high-throughput SNP-typing methods are expected to cover about 80% of the human genome in linkage disequilibrium.[4] A number of large-scale genome-wide cohort studies and case–control studies, such as seven common disease GWAS by the Wellcome Trust Case Control Consortium (WTCCC, 2007), have been planned, and some of them are underway. So far, more than 100 loci of disease-related/causing candidates for about 40 common diseases and traits have been identified,[5] and some loci have led to new insights into pathophysiology and etiological pathways. Because GWAS yields large amounts of raw data and analysis results, the management of GWAS data has become a matter of serious concern. Furthermore, more and more grant-funding agencies, journal editors and research communities are beginning to require the disclosure of GWAS data. Disclosure and data sharing of GWAS data will primarily lead to the following three possibilities: (1) meta-analysis using data sets produced in multiple studies to find novel disease-related SNP candidates; (2) re-use of GWAS data combined with other experimental data, including pathway data and expression data, to deepen the exploration of each disease; and (3) development of methods to analyze and compute genetic statistics. In the case of meta-analysis in particular, the use of raw data is indispensable for quality control and for consideration of population structures. Some studies have successfully found additional disease-related SNP candidates on the basis of meta-analysis.[6,7]

The National Center for Biotechnology Information launched the database (DB) of Genotype and Phenotype in the fall of 2006 as a centralized GWAS system to archive and distribute GWAS data. Currently, results funded by the Genetic Association Information Network and voluntarily submitted data have been accumulated. The European Genotype Archive was created in the spring of 2008 as a repository system for phenotype–genotype relationships, and results primarily from WTCCC have been accumulated and redistributed. To achieve a continuous and intensive management of GWAS data and data sharing among researchers, we established a new DB that is publicly available. This DB is expected to have an essential role in providing easily accessible GWAS data to researchers in various biomedical fields. Some disease-related SNPs are assumed to be buried because of their insufficient *P*-values caused by an insufficient number of case–control samples. It is possible that these SNPs will be revealed by combining the GWAS analysis results with other data possessed by users.

In this paper, we introduce the GWAS DB.

[1]Central Research Laboratory, Hitachi Ltd, Tokyo, Japan; [2]Department of Human Genetics, Graduate School of Medicine, University of Tokyo, Tokyo, Japan; [3]Department of Molecular Life Science and Molecular Medicine, Tokai University School of Medicine, Tokyo, Japan and [4]Department of Neurology, Graduate School of Medicine, University of Tokyo, Tokyo, Japan
Correspondence: Dr A Koike, Central Research Laboratory, Hitachi Ltd, 1-280 Higashi-koigakubo Kokubunji, Tokyo, Japan.
E-mail: asako.koike.ea@hitachi.com
Received 3 June 2009; accepted 27 June 2009; published online 24 July 2009

MATERIALS AND METHODS

Database structure

The DB system consists of an internal GWAS DB and a public GWAS DB. For a maximum of 1 year, or until the acceptance of publication, submitted data are stored in the internal GWAS DB and can be accessed only by the research team that submitted the data for greater convenience in data sharing among research team members living in various locations. Currently, the DB systems are implemented using mysql version 5.0 (http://dev.mysql.com/downloads/mysql/5.0.html), and some of the statistical analysis results are also accumulated in a distributed annotation system (DAS) server. A schematic drawing of the GWAS DB is shown in Figure 1.

In this DB, three types of data access, namely, (1) public access, (2) authorized access accompanied by a data use application, and (3) authorized access accompanied by a data use application and its review by a data access committee, are possible. Principally, frequency data of genotypes and alleles and statistical analysis results can be accessed freely. However, automatic access and frequent access are restricted to prevent the release of frequency data of genome-wide genotypes and alleles, as such a large volume of genotype/allele data leads to the specification of whether the given genome is contained in the case or in the control group, as reported previously.[8] These genome-wide frequency data can be obtained by submitting a data use application to the data access committee. For the use of genotype or raw data, an application that

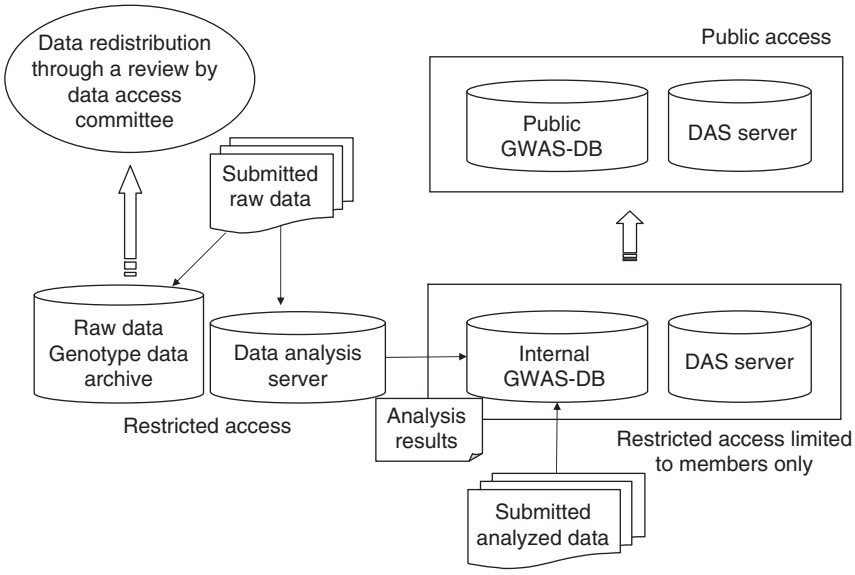

Figure 1 Schematic drawing of genome-wide association study (GWAS) database (DB) systems.

Table 1 Summary of database contents

Contents	*Data sources*
Statistics	
Frequencies of genotypes, alleles and haplotypes	
Statistical genetic analysis	
P-values and odds ratios on genotypic model and allelic model	
P-values and odds ratios on trend model, additive model and recessive model	
Permutation test results	
Bonferroni's corrections and false discovery rate for multiple testing using	
Akaike information criterion	
Hardy–Weinberg equilibrium test	
Haplotype-based χ^2-test	
Epistasis	
Linkage disequilibrium parameters (r^2, D', Lod)	
Other data	
mRNA, amino-acid sequence of each gene	NCBI (http://www.ncbi.nlm.nih.gov/)
mRNA, genome-mapped position	UCSC Hg. 18 (http://hgdownload.cse.ucsc.edu/)
SNP position and SNP kind (cSNP, sSNP, rSNP and so on)	NCBI (http://www.ncbi.nlm.nih.gov/)
OMIM	NCBI (http://www.ncbi.nlm.nih.gov/)
Copy number variation	DGV (http://projects.tcag.ca/variation/)
Gene function	Gene ontology (http://www.geneontology.org/)
Microsatellite polymorphism	UCSC (http://hgdownload.cse.ucsc.edu/)
Manually curated disease-related mutation information	

describes the research purpose and lists the research team members must be submitted to the data access committee. The data access committee deliberates on whether the applicant's research purpose meets the content of the consent form. Only applicants approved by the review committee can use individual genotype data and raw data in accordance with the data handling security rules required by the data access committee and following data use restrictions on the basis of informed consent.

Individual data and raw data are accumulated in the server in a secured computer environment that is different from the public DB server. Only authorized persons can access this server.

Data submission

In principal, both analysis results and unanalyzed data can be submitted. When data have already been analyzed, the analyzed data are accumulated in this DB, along with a detailed description of the analysis protocols. When data have not been analyzed yet, they are analyzed in our site, and the results are accumulated in this DB. When raw data are redistributable under certain conditions, they are also submitted with the contents of the consent form. All data must be submitted with documents explaining the design of the study, as well as ethical consideration.

Data cleaning for quality control

When data are submitted as individual data without analysis results, they are analyzed as follows: (1) SNPs with a call rate $<95\%$ and samples with a call rate $<95\%$ are removed. (2) SNPs, the Hardy–Weinberg equilibrium test result of which in a control group is less than 0.001 or the minor allele frequency of which is less than 0.05, are removed. (3) The principle component analysis (PCA) of these case–control data, along with HapMap data, is carried out using EIGENSTRAT[9] or other programs so that sample outliers and samples with a possible ethnic mixture or a different ethnicity are removed on the basis of the PCA result. Sample outliers in the plot of heterozygosity versus call rate are also removed. The quantile–quantile plot based on the allelic model is calculated and checked. When only genotype frequency data are submitted, PCA and heterozygosity checks are skipped,

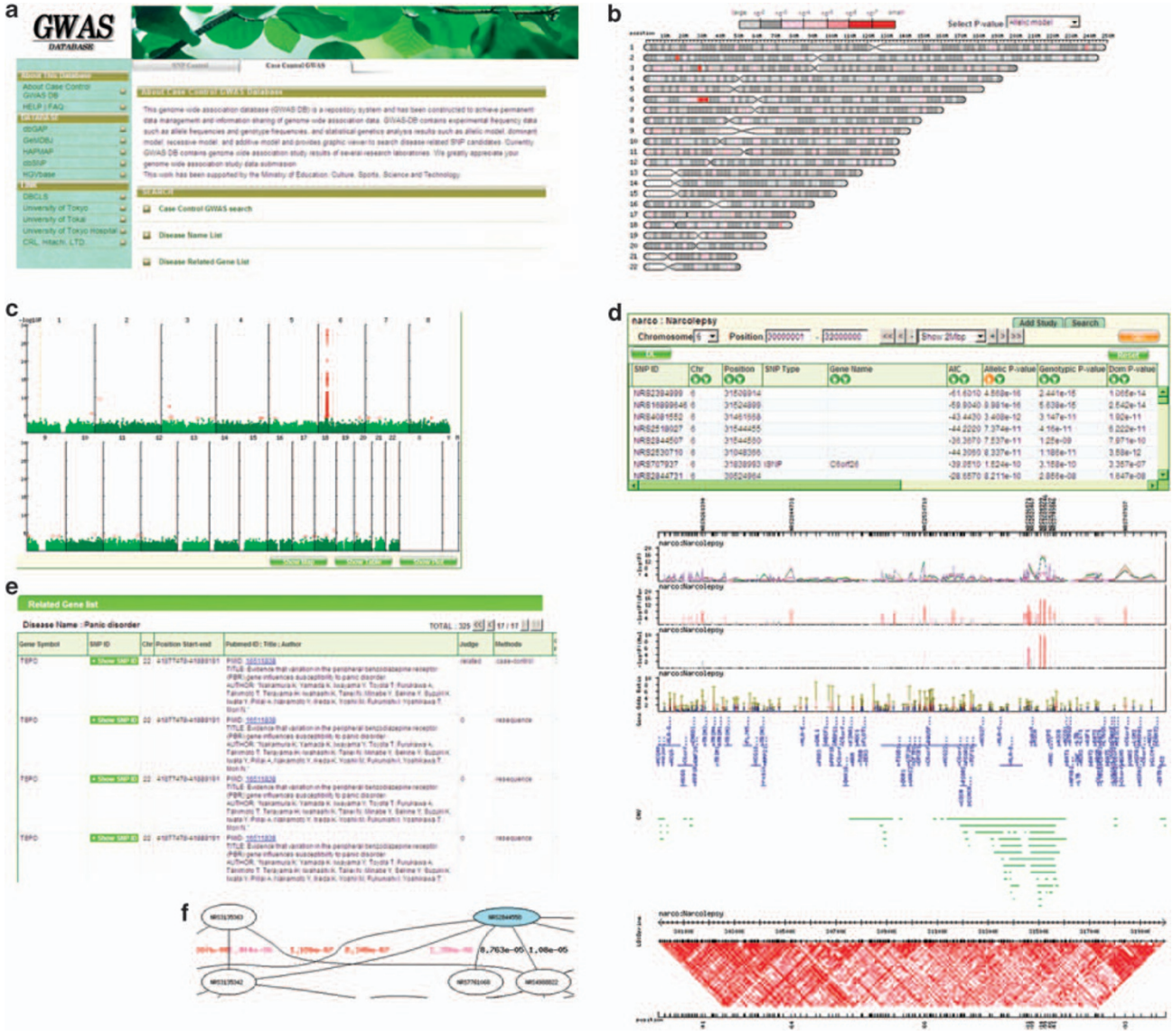

Figure 2 Snapshots of the genome-wide association study (GWAS) database. (**a**) Top page, (**b**) bird's-eye view, (**c**) Manhattan plot, (**d**) region table and graph, (**e**) disease-related gene/single-nucleotide polymorphism (SNP) lists (public data) and (**f**) SNP network based on epistasis.

as they require individual data. The cleaning results are linked from 'study details' on the web.

Data analysis

Standard statistical genetic analyses are performed by plink[10] and Haploview.[11] Additional analyses such as the Akaike information criterion, epistasis and more complicated ones (for example, genetic analysis considering potential case samples existing in the control samples, which sometimes becomes a concern for diseases that develop in old age) are calculated by internally developed programs. The major statistics include *P*-values based on an allelic model, genotypic model, trend model, dominant model, recessive model and permutation test results of these models, and Bonferroni's correction and false discovery rate for multiple testing. These methods are also shown in 'study details.' When submitted data consist of only genotype frequency data, the genome-wide permutation test is skipped.

Database contents and utility

The DB contents (as of April 2009) are summarized in Table 1.

User data other than GWAS data, such as expression data and epigenetic data, are also accumulated and can be displayed on the graph. Although clinical data are not currently accumulated in the DB, they can be added if submitted. Major tables are summarized in Supplementary Table 1.

A snapshot of the GWAS DB is shown in Figure 2. Figure 2a shows the top page of the GWAS DB. When the 'SNP control' tab is selected, the interface jumps to the SNP control DB, which is affiliated to the GWAS DB and contains allelic frequencies, genotypic frequencies, Hardy–Weinberg equilibrium tests and estimated haplotype frequencies of Japanese control samples. Bird's-eye view (Figure 2b) and Manhattan plot (Figure 2c) are provided to draw *P*-values of each model. A genome region can be selected from both (Figures 2b and c), and the results of statistical genetic analysis along with other information such as exon–intron information and copy number variations (CNVs) can be displayed in tables and graphs to facilitate the identification of disease-related SNPs, as shown in Figure 2d. Furthermore, comparisons among various study results obtained by different institutions and/or different platforms can be carried out easily by plotting their graphs on the web (using the 'add study' function in Figure 2d). When the published disease-related gene or SNP is registered as shown in Figure 2e, data are plotted as a known disease-related gene/SNP in the graph (Figure 2d). Epistasis data are also accumulated and drawn as a network graph using Graphviz (http://www.graphviz.org/), as shown in (Figure 2f). Data can be searched by SNP ID (dbSNP ID #rs, affymetrix SNP ID and so on), gene name, disease name and so on. The study design and analysis protocols can also be browsed.

Statistical results are also accumulated on a DAS server, and they can be browsed using the Gmod Gbrowse (http://gmod.org/wiki/Main_Page)-based browser (http://gwas.lifesciencedb.jp/cgi-bin/gbrowse/snpdb/). Furthermore, as a function of the DAS server, data on other DAS servers such as Ensemble can be called up. This function is useful to superimpose data from other DBs onto GWAS data. The GWAS DB is designed to be user friendly for researchers unfamiliar with GWAS to promote disease-related studies.

Further development

A recent topic of interest is genome-wide association analysis coupled with other data such as pathway data[12] to compensate for the low statistical power in disease-associated candidate SNPs. The function to browse or calculate SNP/SNP pair *P*-values on the basis of the GWAS result, along with other data, will be added to this DB to facilitate the generation and understanding of user hypotheses.

The relationships between CNVs and diseases have begun to emerge in recent studies.[13] Although concerns remain about the quality of detected CNVs, genomic locations and frequencies of CNV regions and their case–control association study results will be incorporated into this DB. Furthermore, in the near feature, new high-throughput techniques such as short-read sequencing will be applied for GWAS, and this DB will be improved to suit the new experimental techniques.

ACKNOWLEDGEMENTS

This work was supported by the contract research fund 'Integrated Database Project' from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

1 Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2 Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
3 The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
4 Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38,** 659–662 (2006).
5 Manolio, T. A., Brooks, L. D. & Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118,** 1590–1605 (2008).
6 Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40,** 638–645 (2008).
7 Houlston, R. S., Webb, E., Broderick, P., Pittman, A. M., Di Bernardo, M. C., Lubbe, S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* **40,** 1426–1435 (2008).
8 Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4,** e000167 (2008).
9 Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
10 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
11 Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21,** 263–265 (2005).
12 Baranzini, S. E., Galwey, N. W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D. *et al.* Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.* **18,** 2078–2090 (2009).
13 McCarroll, S. A. Extending genome-wide association studies to copy-number variation. *Hum. Mol. Genet.* **17** (R2), R135–R142 (2008).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)