Published online 19 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W17–W22***
doi:10.1093/nar/gkp355

Hyperlink Management System and ID Converter System: enabling maintenance-free hyperlinks among major biological databases

Tadashi Imanishi[1,*] and Hajime Nakaoka[1,2,3]

[1]Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo 135-0064, [2]Japan Biological Informatics Consortium, Time 24 Bldg. 10F, 2-45 Aomi, Koto-ku, Tokyo 135-8073 and [3]C's Lab Co., Ltd., Kinyo-Kanda Bldg. 7F, 1-3 Kanda-Tomiyama-cho, Chiyoda-ku, Tokyo 101-0043, Japan

Received February 1, 2009; Revised April 16, 2009; Accepted April 23, 2009

ABSTRACT

Hyperlink Management System (HMS) is a system for automatically updating and maintaining hyperlinks among major public databases in the field of life science. We daily create corresponding tables of data IDs of major databases for human genes and proteins, and provide a CGI-program that returns correct and up-to-date URLs for showing data of various databases that correspond to user-specified IDs. The HMS can deal with various IDs: accession numbers of International Nucleotide Sequence Databases, HUGO Gene Symbols and IDs of UniProt, PDB, H-InvDB and others, and it can return URLs of various databases: H-InvDB, HUGO Gene Nomenclature Committee Database, NCBI Entrez Gene, UniProt, PDB and others. For example, 23 297 pages of Locus view of H-InvDB are reachable by using HUGO Gene Symbols through the HMS. Not only the CGI-program, the HMS provides a Web page for finding and opening URLs of these databases. Although hyperlinking is an effective way of relating biological data among different databases, updating hyperlinks has been a laborious work. The HMS fully automates the job, enabling maintenance-free hyperlinks. We also developed the ID Converter System (ICS) for simply converting data IDs by using corresponding tables in the HMS. The HMS and ICS are freely available at http://biodb.jp/.

INTRODUCTION

There are currently 1170 biological databases in the world (1) that are being utilized for various researches and developments. It is now possible for biologists to try to make discoveries by combining and linking data of different types obtained from multiple databases. For example, by combining protein sequences and protein structure data that are stored separately in major databases, biologists can study the basic rules of protein 3D structure such as how amino-acid sequences organize themselves into specific and stable 3D structures. In this way, linking specific data stored in multiple databases will lead to the organization of biological knowledge and will provide useful resources for novel researches.

Finding corresponding data in different databases and setting hyperlinks among them is an effective way of integrating information that are stored in independently operated databases in different organizations. The maintenance of the hyperlinks among many databases, however, requires a huge cost, because the number of pairs of databases is generally very large. For example, possible ways of hyperlinks among N related databases will be $(N-1)N$, if we consider the directions of hyperlinks. Suppose if one of the N databases is updated, $2(N-1)$ ways of hyperlinks need to be updated. If each of N databases is periodically updated in every 3 months, at most $8(N-1)N$ ways of hyperlinks need to be updated in a year, which will be a laborious work. We thus need a novel system to automate the maintenance of hyperlinks.

Here, we developed a Web server for automatically managing and updating hyperlinks that can connect

*To whom correspondence should be addressed. Tel: +81 3 3599 8800; Fax: +81 3 3599 8801; Email: t.imanishi@aist.go.jp
This website is free and open to all users and there is no login requirement. The website address is 'http://biodb.jp/'. The website was officially made public on April 2007, and has been running since then. The Hyperlink Management System has been used by the HUGO Gene Nomenclature Committee (HGNC) Database, GeneCards, H-InvDB and other public databases to maintain hyperlinks among these databases. The ID Converter System has been used by hundreds of users since we opened it in March 2008.

molecular information in multiple databases about human genes and proteins. Furthermore, we developed a Web server for converting data IDs from one type to another. These Web servers will be useful for both database developers and biologists. The reason for it is that, for database developers, these systems can realize automated, work-saving and efficient management of databases, and can reduce the large cost of maintaining biological databases. Also, for biologists, especially those who are unfamiliar with the data IDs of major biological databases, our newly developed Web servers will be of help in carrying out efficient analyses of 'big data' that are being produced massively in laboratories (2). We hope that many database managers and researchers will find our Web servers useful.

WEB SERVER FEATURES

Usage of the Hyperlink Management System (HMS) CGI-program

HMS is a tool for automatically updating and maintaining hyperlinks between databases in the field of life science. The HMS automatically downloads the lists of data IDs for human genes and proteins and their relations from major biological databases everyday, and produces large corresponding tables of all data IDs of these databases (Tables 1 and 2). Then, using the corresponding tables of data IDs, the HMS shows the most up-to-date URLs for data in a database that are corresponding to a user-specified ID of another database. The URL is easily obtained by running a CGI-program of the HMS.

Suppose that the manager of database *A* wishes to create hyperlinks from database *A* to database *B*. Usually in this case, the manager uses data IDs of database *B* in the URLs of hyperlinks. However, if the manager uses the HMS, he can use data IDs of database *A* to create hyperlinks to database *B*. This is very useful because we can easily create hyperlinks without knowing data IDs of target databases. We can therefore drastically reduce the cost of creating and maintaining hyperlinks among databases.

Currently, the HMS collects 17 types of data IDs for human genes and proteins (Table 2). Among the 17 types of data IDs, we defined accession numbers of International Nucleotide Sequence Database (INSD; DDBJ/EMBL/GenBank), HUGO Gene Symbols and UniProt IDs as 'common IDs' that are used as key IDs in creating corresponding tables. We collect corresponding tables between a common ID and an other ID from each of source databases (Figure 1). We also obtain a corresponding table between accession numbers and HUGO Gene Symbols from NCBI and a corresponding table between accession numbers and UniProt IDs from UniProt. Using these tables, we create a large corresponding table that shows the relationships among 17 types of data IDs. The common IDs define the route of ID conversion (as shown in Figure 1). For example, when converting HIX of H-InvDB into PDB IDs, we first convert HIX into UniProt IDs and then convert it into PDB IDs.

The HMS directly opens a Web page of the target database if there is only one data ID corresponding to the source ID. However, in converting data IDs, an ID does not always correspond to one ID of another type. If there

Table 1. Databases and viewers handled by the HMS

Database name	Viewer name	Format[a]
H-InvDB (3,4)	Transcript view	TRANSCRIPTVIEW
	Locus view	LOCUSVIEW
	G-integra	GINTEGRA
	PPI view	PPI
INSD (DDBJ/EMBL/GenBank) (5–7)	NCBI Nucleotide	NUCLEOTIDE
NCBI Entrez Gene (8)	Entrez Gene	ENTREZGENE
Genome Medicine Database of Japan (GeMDBJ) (9)	GeMDBJ	GEMDBJ
HUGO Gene Nomenclature Committee (HGNC) Database (10)	HGNC	HUGO
MutationView (11)	MutationView	MUTATIONVIEW
H-GOLD (12)	H-GOLD	HGOLD
PDB (13)	PDB	PDB
UniProt (14)	UniProtKB	UNIPROT
Ensembl (15)	Gene Summary	ENSG
	Transcript Summary	ENST
GlycoGene DataBase (GGDB) (16)	GGDB	GGDB
fRNAdb (17)	fRNAdb	FRNADB
Human Gene and Protein Database (HGPD) (18)	HGPD	HGPD

[a]Parameters of database names used in the HMS CGI-program, ICS and Web service.

Table 2. Data IDs used in the HMS

Source database	Data ID	Format[a]	Count[b]
H-InvDB	H-Inv transcript ID (HIT)	HIT_ID	219 765
	H-Inv cluster ID (HIX)	HIX_ID	43 159
	H-Inv protein ID (HIP)	HIP_ID	133 629
INSD	Accession Number	ACC_ID	222 136
NCBI	OMIM ID	OMIM_ID	16 253
	RefSeq ID	REF_SEQ	48 671
	Entrez Gene ID	GENE_ID	40 235
GeMDBJ	dbSNP rs#	JSNP_ID	76 024
HUGO	HUGO Gene Symbol	GENE_SYMBOL	26 421
H-GOLD	H-GOLD Marker Name	MS_ID	494
PDB	PDB ID	PDB_ID	50 958
UniProt	UniProt ID	UNIPROT_ID	84 422
Ensembl	Ensembl Transcript ID	ENST	63 280
	Ensembl Gene ID	ENSG	37 435
fRNAdb	fRNAdb ID	FR_ID	41 501
HGPD	cDNA clone ID	PDB_ID	21 340
	FLJ cDNA clones	FLJ_ID	21 340

[a]Parameters of ID types used in the HMS CGI-program, ICS and Web service.
[b]Number of data as of 23 January 2009.

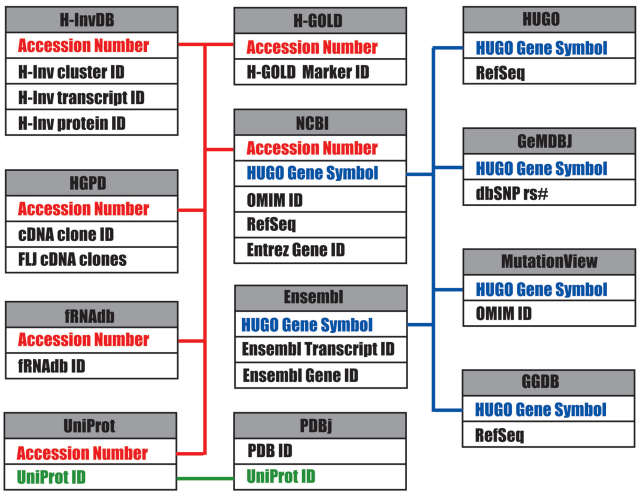

Figure 1. Relationships among data IDs obtained from 13 databases. Here, INSD is represented by NCBI. For all these IDs, the HMS and the ID Converter System are capable of setting hyperlinks and ID conversion. 'Common IDs' that were used as key IDs in ID conversion are shown in red, blue or green. Lines connecting databases represent the paths of ID conversion. For example, when converting HUGO Gene Symbols into PDB IDs, the system first converts HUGO Gene Symbols into accession numbers according to a corresponding table obtained from NCBI, and then converts accession numbers into UniProt IDs according to a corresponding table obtained from UniProt, and then converts UniProt IDs into PDB IDs according to a corresponding table obtained from PDBj. Live lists of data IDs and their corresponding tables are collected daily from each of these databases.

are more than one IDs corresponding to the source ID, the HMS shows the list of all corresponding data IDs with appropriate hyperlinks, from which users can choose and open the data of interest.

In order to create hyperlinks using the HMS CGI-program, use the following URL as a target URL.

http://biodb.jp/hfs.cgi?id=[ID]&type=[ID Type]&db=[Database Name]

Here, [ID] is a data ID of the database from which hyperlink is set, [ID Type] is a kind of the data ID and [Database Name] is the name of database to which hyperlink is set. Details of parameter settings are shown in Tables 1 and 2 and also in the online HELP pages of the HMS. For example, using BC053 657, an accession number of INSD, you can create a hyperlink toward the corresponding 'Transcript View' of H-InvDB in the following way.

BC053 657

This CGI-program is freely available to all users in the world. For database managers, if their databases have any one of those data IDs shown in Table 2, it is possible to create hyperlinks from the database to all the databases shown in Table 1 through the HMS. Because the hyperlinks of the HMS are automatically updated everyday, the connection to the latest data is always guaranteed.

Usage of the HMS Web Server

The HMS CGI-program is supposed to be used for hyperlinking among databases. However, there are situations that we want to query with a small number of data IDs. We thus developed a Web server of the same function as the CGI-program (Figure 2). Here, by utilizing the same corresponding tables of data IDs as those of the CGI version, users can open Web pages of a database corresponding to a specified data ID of another database. The usage of the HMS Web server is quite simple and requires no complicated explanations. For example, if you enter an accession number of INSD in the 'Source ID' and push the 'Search' button, you can open the 'Locus view' of H-InvDB. In the textbox, users can enter multiple data IDs delimited by a space, tab or comma. In this case, the result page shows a list of multiple URLs with hyperlinks corresponding to each data.

Usage of the ID Converter System (ICS)

ICS is a tool for converting data IDs used in a database into other, corresponding IDs used in other databases (Figure 3). All the data IDs used in the HMS can be dealt with. Multiple IDs can be converted at a time, if users enter multiple data IDs delimited by a space, tab or comma. When a data ID can be converted into multiple corresponding IDs, the ICS shows all the corresponding IDs. Users can also specify a file on their client PC, and convert a list of data IDs in it. Furthermore, the ICS is accessible by computer programs using the Web service, as will be described later.

Web service and other functions

For those who wish to analyze massive data by computers, the HMS and ICS accept queries through the Web service, which makes it possible to call functions of these systems through computer programs. The details of commands for the Web service will be shown if you click the 'Web service' button on the upper right corner of the Web pages (Figures 2 and 3). On these pages, there are some sample programs for the Web service. Furthermore, we prepared statistics pages of the HSM and ICS. If you click the 'Update Info' button on the upper right corner of the top pages (Figures 2 and 3), the number of data IDs downloaded from each database everyday will be shown. If you click the 'Release Info' button, the number of pages of each database that are hyperlinked through various data IDs will be shown. Finally, there are help pages that explain how to use the HMS and ICS.

DISCUSSION

As has been stated earlier, there are currently 1170 databases in the field of life sciences that are listed in the online Molecular Biology Database Collection of Nucleic Acids

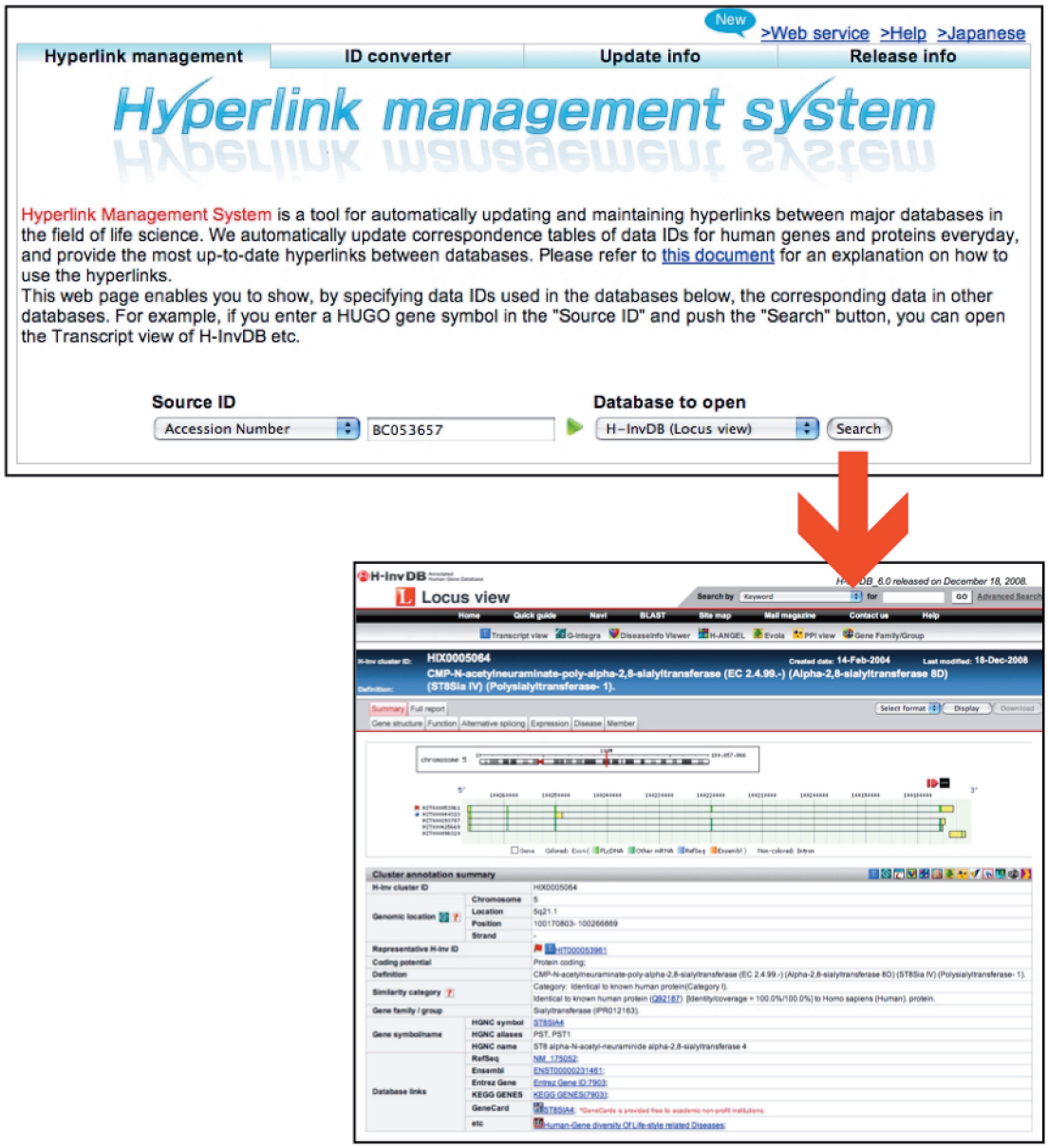

Figure 2. Usage of the HMS on the Web. Input source IDs in the textbox, and select the type of IDs. Then, select database to open, and push the 'Search' button. Then, the HMS searches for IDs that correspond to the source IDs and opens pages of destination databases. If there are two or more corresponding data, the HMS shows the list of IDs with hyperlinks. This figure shows an example of opening Locus view of H-InvDB with an accession number BC053657.

Research (1). The number of databases is increasing every year. Because biologists usually wish to combine a certain type of data with another, integration of these various and huge databases will be of great significance. In principle, there are two possible ways of integrating databases. One is to collect all the databases at one place and merge them, while the other is to combine the databases virtually by hyperlinking corresponding data among them. The former method requires an enormous amount of disk space of computers and a large number of people to maintain and update the databases. In other words, the cost of the former method is too large to implement. In fact, there was an attempt to construct a huge integrated database of many biological data by the former method, which failed in a few years probably due to the high running cost. The latter method, on the other hand, can be realized relatively easily if we adopt the automated hyperlinking that we introduced in this article. The HMS will be a fundamental platform to virtually integrate databases that are operated independently by different organizations. Furthermore, the HMS will make it possible for us to construct integrated databases in a sustainable way.

The ICS is also a quite useful tool. Because major databases such as INSD and UniProt update their 'live lists' of

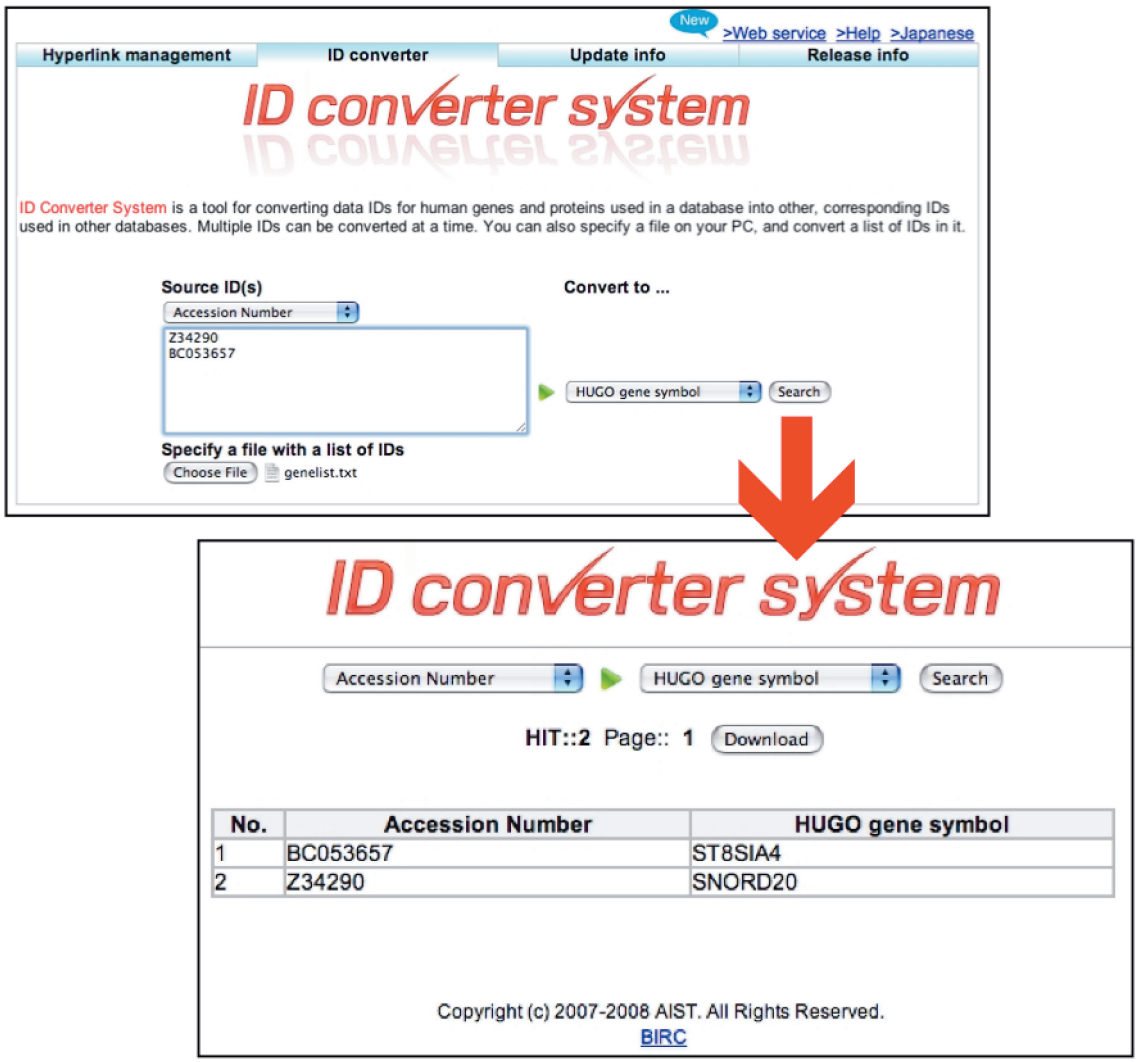

Figure 3. Usage of the ID Converter System on the Web. Input source IDs in the textbox, and select the type of IDs. You may also specify a text-file with a list of source IDs. Then, select the type of data ID to convert, and push the 'Search' button. Then, the system searches for IDs that correspond to source IDs, and shows the list of them. The list can be downloaded by pushing the 'Download' button. This figure shows an example of converting two accession numbers, BC053657 and Z34290, into HUGO Gene Symbols.

data IDs on a daily basis, the role of ICS to provide users with the up-to-date information of ID conversion is very important. The usage of the ICS is very simple, and anyone can make full use of it easily. For example, even when one has a list of HUGO gene symbols for 100 genes, he/she can easily find corresponding data in H-InvDB, UniProt, PDB and other databases through the ICS. Advantages of the ICS are not limited to this. The most important feature of the ICS is that ID conversion can be done between two types of data IDs that are not directly related. The ICS can convert from and to all data IDs shown in Figure 1, including those that are not directly connected. For example, the ICS can directly convert accession numbers of ISND into PDB IDs. We thus think that the ICS will be a useful Web server for many researchers.

We plan to add more and more databases into the HMS and ICS in the future. It is easy to add new databases as destination of HMS hyperlinks if the database uses either accession numbers of INSD, UniProt IDs or PDB IDs. We welcome any proposals about incorporating additional databases into the HMS. Furthermore, we have been dealing with databases about human molecular data, but we plan to extend the HMS to molecular data of other species in near future. The ICS will be also extended to incorporate other IDs that are frequently used by biologists. For example, we will incorporate IDs of literature databases, which will enable combining biological knowledge in literature with molecular data. We will extend the ICS according to the requests from users.

The HMS and ICS will be of considerable use as an automated tool for database managers and as a new

indispensable tool for biologists. We may be able to extend these systems to incorporate keyword searches that are frequently used in many databases. By indexing general and technical terms by assigning them some IDs, the HMS and ICS will be able to deal with keyword searches. Such an extension will further increase the value of these systems.

ACKNOWLEDGEMENTS

We thank managers of collaborating databases for kindly providing lists of data IDs. We also thank research staffs in BIRC, AIST for helpful suggestions and technical support.

FUNDING

Integrated Database Project of the Ministry of Economy, Trade and Industry of Japan; research grant for promoting science and technology from Japan Science and Technology Agency. Funding for open access charge: National Institute of Advanced Industrial Science and Technology.

Conflict of interest statement. None declared.

REFERENCES

1. Galperin,M.Y. and Cochrane,G.R. (2009) Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. *Nucleic Acids Res.*, **37**, D1–D4.
2. Howe,D., Costanzo,M., Fey,P., Gojobori,T., Hannick,L., Hide,W., Hill,D.P., Kania,R., Schaeffer,M., St Pierre,S. *et al.* (2008) Big data: the future of biocuration. *Nature*, **455**, 47–50.
3. 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**, 856–875.
4. Yamasaki,C., Murakami,K., Fujii,Y., Sato,Y., Harada,E., Takeda,J., Taniya,T., Sakate,R., Kikugawa,S., Shimada,M. *et al.* (2008) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res.*, **36**, D793–D799.
5. Sugawara,H., Ikeo,K., Fukuchi,S., Gojobori,T. and Tateno,Y. (2009) DDBJ dealing with mass data produced by the second generation sequencer. *Nucleic Acids Res.*, **37**, D16–D18.
6. Cochrane,G., Akhtar,R., Bonfield,J., Bower,L., Demiralp,F., Faruque,N., Gibson,R., Hoad,G., Hubbard,T., Hunter,C. *et al.* (2009) Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res.*, **37**, D19–D25.
7. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
8. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
9. Yoshida,T. and Yoshimura,K. (2003) Outline of disease gene hunting approaches in the Millennium Genome Project of Japan. *Proc. Japan Acad.*, **79**, 34–50.
10. Bruford,E.A., Lush,M.J., Wright,M.W., Sneddon,T.P., Povey,S. and Birney,E. (2008) The HGNC database in 2008: a resource for the human genome. *Nucleic Acids Res.*, **36**, D445–D448.
11. Minoshima,S., Mitsuyama,S., Ohtsubo,M., Kawamura,T., Ito,S., Shibamoto,S., Ito,F. and Shimizu,N. (2001) The KMDB/MutationView: a mutation database for human disease genes. *Nucleic Acids Res.*, **29**, 327–328.
12. Tamiya,G., Shinya,M., Imanishi,T., Ikuta,T., Makino,S., Okamoto,K., Furugaki,K., Matsumoto,T., Mano,S., Ando,S. *et al.* (2005) Whole genome association study of rheumatoid arthritis using 27 039 microsatellites. *Hum. Mol. Genet.*, **14**, 2305–2321.
13. Henrick,K., Feng,Z., Bluhm,W.F., Dimitropoulos,D., Doreleijers,J.F., Dutta,S., Flippen-Anderson,J.L., Ionides,J., Kamada,C., Krissinel,E. *et al.* (2008) Remediation of the protein data bank archive. *Nucleic Acids Res.*, **36**, D426–D433.
14. The UniProt Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
15. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
16. Narimatsu,H. (2004) Construction of a human glycogene library and comprehensive functional analysis. *Glycoconj. J.*, **21**, 17–24.
17. Mituyama,T., Yamada,K., Hattori,E., Okida,H., Ono,Y., Terai,G., Yoshizawa,A., Komori,T. and Asai,K. (2009) The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Res.*, **37**, D89–D92.
18. Maruyama,Y., Wakamatsu,A., Kawamura,Y., Kimura,K., Yamamoto,J., Nishikawa,T., Kisu,Y., Sugano,S., Goshima,N., Isogai,T. *et al.* (2009) Human Gene and Protein Database (HGPD): a novel database presenting a large quantity of experiment-based results in human proteomics. *Nucleic Acids Res.*, **37**, D762–D766.