

Gene Nomenclature Committee (HGNC) and Rat Genome Database (RGD) with nomenclature activities for genes, alleles and strains for each species (2–4). Data from the National Center for Biotechnology Information (NCBI) and Ensembl are also broadly used across species (5,6). The Open Biomedical Ontology (OBO) Consortium, an umbrella body for the developers of life-science ontologies, also provides ontologies developed with the aim of comprehensive annotation of biological information (7,8).

In the mouse genetical research community, these issues have been discussed by international consortia. The Mouse Phenotype Database Integration Consortium (InterPhenome) (<http://www.interphenome.org/>) and the Coordination and Sustainability of International Mouse Informatics Resources (CASIMIR) (<http://www.casimir.org.uk/>) have discussed broad issues regarding the integration, coordination, interoperability and sustainability of databases, such as methodologies to integrate phenotype information, the association of phenotype with human disease, models for long-term and financial sustainability for databases and legal issues of data accessibility (9–11). A complete solution to satisfy these multiple and broad requirements at once is desired to ensure the sustainability of databases.

One effective way to reduce the management cost of databases is to share common fundamental infrastructures such as the hardware and application software used in their implementations. Recently, such common operations have been effectively implemented through ‘cloud computing’, which is a type of internet-based computing whereby shared resources, software and information are provided on demand. Cloud computing is often economically beneficial for the facility in terms of the running costs of space, electricity, cooling and staff support (12). If data are properly and continuously managed and integrated with the public data records that are regarded as the *de facto* standard in the biomedical community, then a common infrastructure could be one of the best ways to achieve cost effectiveness and advanced usability. On the other hand, the ‘semantic web’ offers a series of methods and technologies to develop extensions of the current World Wide Web (WWW) in which information is given well-defined meanings and integrated (13). These technologies include the Resource Description Framework (RDF), a variety of data interchange formats (e.g. RDF/XML, N3, Turtle and N-Triples), and notations such as the RDF Schema (RDFS) and the Web Ontology Language (OWL), all of which are intended to provide a formal description of concepts, terms and relationships within a given knowledge domain. The semantic web is regarded as an integrator across different content and information applications and systems and provides mechanisms for the realization of a common information system. It is also useful for the dissemination of data, providing a standardized framework to describe metadata recommended by the WWW consortium that aids the automated (and also manual) processing of disseminated data to derive meaning from the data. The dissemination of data with standardized metadata risks the extinction of the data and creates the

opportunity to promote the discovery of new knowledge. Consequently, the semantic web seems to be suitable as a fundamental technology to implement the common infrastructure.

In this study, we developed a new database, the RIKEN integrated database of mammals, as an official undertaking in RIKEN to integrate heterogeneous mammal-related data in multiple individual databases. This database was constructed on the Scientists’ Networking System (SciNetS: [http://www.riken.jp/eng/r-world/info/release/press/2009/090331\\_2/](http://www.riken.jp/eng/r-world/info/release/press/2009/090331_2/)), a general fundamental system that applies the semantic web technology to provide massive data management, supported by Japan’s national database integration project. In this system, we achieved the top-level ontology-based re-organization of imported data to integrate the typical and instructive knowledge with individual data records. The RIKEN integrated database of mammals is complementary to the original databases. For example, the FANTOM web resource aims to present data on the dynamic behavior of transcription and its regulation in the expanding fields of the transcriptome, epigenome and transcriptional networks (14,15). By contrast, this integrated database attaches greater importance to the standardization of data for better distribution, metadata-level integration and cross-database retrieval.

## DATABASES TO BE INTEGRATED

In RIKEN, there are a number of databases related to mammalian research resources. In the primary development of the integrated database, we integrated six database projects: the Functional Annotation of the Mammalian Genome 4 (FANTOM 4: <http://fantom.gsc.riken.jp/>) (14–16), the RIKEN Cerebellar Development Transcriptome Database (CDT-DB: <http://www.cdt.db.brain.riken.jp/CDT/Top.jsp>) (17,18), the resource database from the RIKEN BioResource Center (BRC) (19–21) including mutant resources produced by the ENU mutagenesis program (22,23) and the Resource of Asian Primary Immunodeficiency Diseases (RAPID) (24), the RIKEN Structural Genomics/Proteomics Initiative (RSGI) and two data repositories for the Reference Database of Immune Cells (RefDIC) (25) and the RIKEN Expression Array Database (READ) (26), all of which are produced from individual research projects in the human and mouse. Each database project has its original data schema to represent a variety of data ranging from research resources, such as biological strains, cell lines and DNA clones, to experimental data, such as gene expression and phenotypic analyses. There are no relationships defined among original data tables, which are described by various data formats such as text, images and movies. However, as is usual for most databases, they are compiled in a main data table to represent the objects of the database and related information (Table 1). In the discussions of InterPhenome and CASIMIR, it was recommended that the equivalences or relationships among records from the MGI database for genes and alleles, the International Mouse Strain

Table 1. Imported databases in RIKEN (as for September 2010)

Database (URL)	Contents	Project URL in SciNetS
FANTOM4 ( <a href="http://fantom.gsc.riken.jp/4/">http://fantom.gsc.riken.jp/4/</a> )	Monitoring of the dynamics of transcription start site (TSS) usage during a time course of monocytic differentiation in the acute myeloid leukemia cell line THP-1.	<a href="http://scinets.org/item/ria187i/">http://scinets.org/item/ria187i/</a>
Bio-resource catalog ( <a href="http://www.brc.riken.jp/">http://www.brc.riken.jp/</a> )	The online catalog of bioresources including mammalian laboratory strains (mouse), cells and DNA clones in the RIKEN BioResource Center (BRC).	<a href="http://scinets.org/item/ria256i/">http://scinets.org/item/ria256i/</a>
RIKEN ENU Mouse Lines ( <a href="http://www.brc.riken.jp/lab/gsc/mouse/">http://www.brc.riken.jp/lab/gsc/mouse/</a> )	Phenotype information of mutant mouse lines generated from large-scale ENU mutagenesis as a resource of the RIKEN BRC.	<a href="http://scinets.org/item/rib190i/">http://scinets.org/item/rib190i/</a>
Pheno-Pub ( <a href="http://www.brc.riken.jp/lab/jmc/mouse_clinic/en/m-strain_en.html">http://www.brc.riken.jp/lab/jmc/mouse_clinic/en/m-strain_en.html</a> )	Phenotype data from the standardized phenotyping platform of the Japan Mouse Clinic (JMC) project in the RIKEN BRC.	<a href="http://scinets.org/item/ria110i/">http://scinets.org/item/ria110i/</a>
Cerebellar Development Transcriptome Database (CDT-DB: <a href="http://www.cdtb.brain.riken.jp/CDT/Top.jsp">http://www.cdtb.brain.riken.jp/CDT/Top.jsp</a> )	The spatio-temporal gene expression profile of the postnatal development of the mouse cerebellum,	<a href="http://scinets.org/item/cria237u1i/">http://scinets.org/item/cria237u1i/</a>
Resource of Asian Primary Immunodeficiency Diseases (RAPID: <a href="http://rapid.rcai.riken.jp/RAPID">http://rapid.rcai.riken.jp/RAPID</a> )	A web-based compendium of molecular alterations in primary immunodeficiency diseases.	<a href="http://scinets.org/item/cria271u1i/">http://scinets.org/item/cria271u1i/</a>
Systems and Structural Biology Center (SSBC) database ( <a href="http://www.rsgi.riken.jp/rsgi_e/index.html">http://www.rsgi.riken.jp/rsgi_e/index.html</a> )	The crystal structures of proteins and the protein-protein interactions in living cells analyzed with the expansion of the genetic code.	<a href="http://scinets.org/item/ria46i/">http://scinets.org/item/ria46i/</a>
Reference Database of Immune Cells (RefDIC: <a href="http://refdic.rcai.riken.jp/welcome.cgi">http://refdic.rcai.riken.jp/welcome.cgi</a> )	An open-access database of quantitative mRNA and protein profiles specifically for immune cells and tissues.	<a href="http://scinets.org/item/crib225s27rib225s7i/">http://scinets.org/item/crib225s27rib225s7i/</a>
RIKEN Expression Array Database (READ: <a href="http://read.gsc.riken.jp/">http://read.gsc.riken.jp/</a> )	An integrated system for microarray data that works like 'glue' in post-sequence and post-hybridization analyses.	<a href="http://scinets.org/item/crib225s27rib225s8i/">http://scinets.org/item/crib225s27rib225s8i/</a>

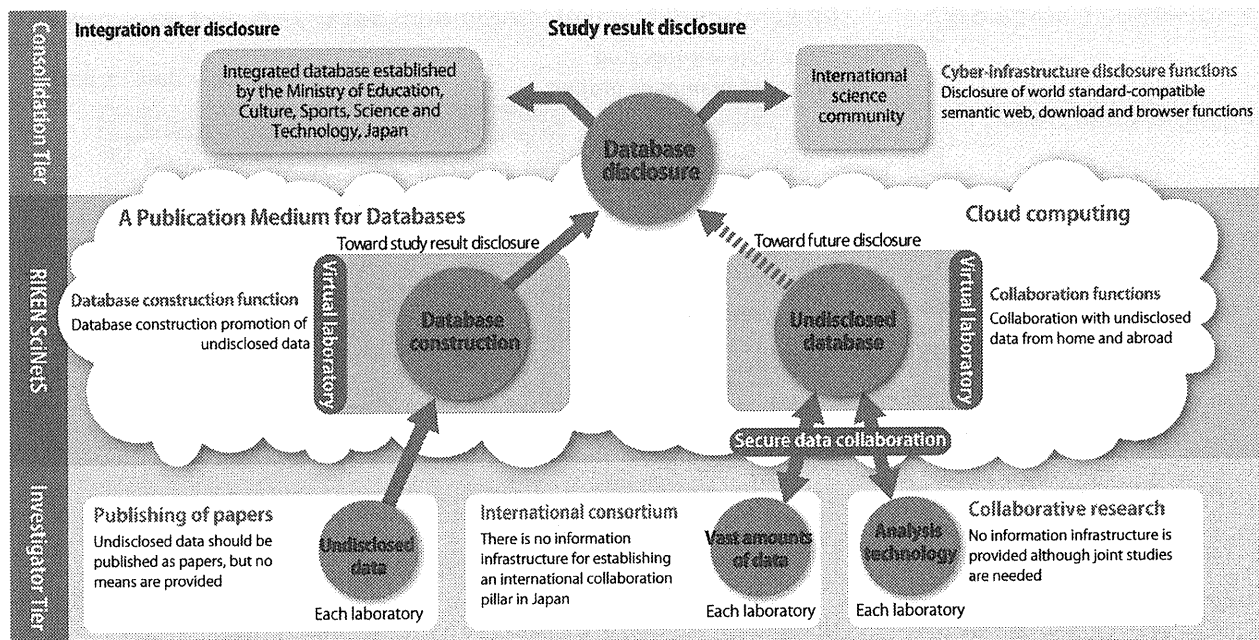
Resource (IMSR) for experimental strain (27) and terms of OBO ontologies be specified. To show the association between the institute's data and the public data broadly used in the research community, we constructed an association between RIKEN's data and public data (Supplementary Table S1).

THE FUNDAMENTALS OF THE INTEGRATED DATABASE: SEMANTIC WEB-BASED CLOUD SYSTEM 'SCINET'S'

We have implemented the integrated database on the data-hosting system, SciNetS, which is a fully web-based common platform that ensures cloud computing in the scientific community on the basis of semantic web technologies (Figure 1). It has multiple features useful for data integration:

- (1) The system is designed to support sharing of academic information with secure, and to handle databases for sharing, collaborating or publication. Database developers can set multiple levels of accessibility within user groups or the public for each 'project' (private workspace). The user can also declare the copyright licenses for their digital

- content with Creative Commons (CC) or GNU to indicate the availability for secondary use.
- (2) In the project, database developers can also design the semantics with elements and the equivalent methodology to RDF and OWL-Full (i.e. the definition of the semantic links between class and subclass, class and instance, property and sub-property and so on) with graphic user interfaces (GUI) for ontology editors such as Protégé (28). The system assigns the Uniform Resource Identifier (URI) to each data element.
- (3) The structured data and metadata can be placed in the public directories of the SciNetS with various standardized data formats, such as RDF, OWL or tab-delimited text file, for downloading or direct connection from external systems and application software such as Protégé.
- (4) The system provides the tracking back function making automatically reverse links for RDF relationships across projects. It ensures automatic integration of distributed effort of annotation and curation.
- (5) The system is designed to handle a large number of databases simultaneously and is scalable for increased data with the distributed processing technologies on databases and query functions (29).



**Figure 1.** Schematic diagram showing the concept of SciNetS. SciNetS provides incubation functions from database construction to the integration of databases in computing clouds or a group of large-scale servers, and discloses databases using interfaces compatible with international standards, thus contributing to the establishment of cyber-infrastructure for integrating worldwide databases [reprinted with the courtesy from Tetsuro Toyoda, 'Synthetic biology—creating biological resources from information resources' RIKEN RESEARCH 5(10) 13–16, 2010 (<http://www.rikenresearch.riken.jp/eng/frontline/6397>)].

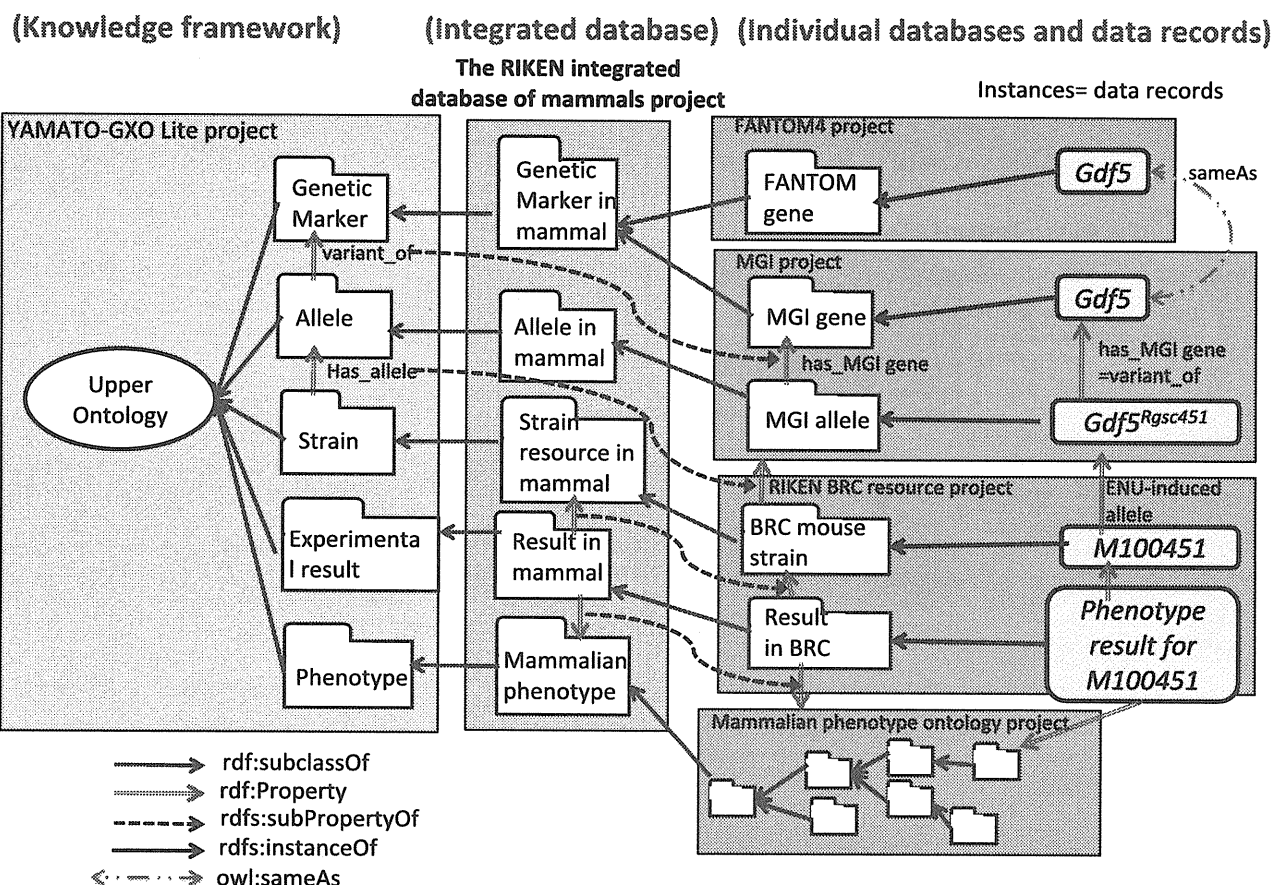
The high-speed retrieval of semantic content is implemented with the General and Rapid Association Study Engine (GRASE), which enables semantic Boolean-based deduction and statistical evaluation of RDF resources (29).

## IMPLEMENTATION OF MAMMALIAN DATA IN SCINET S

The overview of the implementation of this database is presented in Figure 2. The mammalian data and public data shown in Table 1 and Supplementary Table S1, respectively, were imported to SciNetS as individual database projects such that their intact data schema were reflected fully or partially. According to the forms of the original data sources, the databases were imported as three distinct types of projects implemented in SciNetS. First, in the database-type project, a replication of the original database elements, the database table and a data record, is represented with a class and an instance, respectively. Second, the ontology-type project is a replication of the ontology with the OWL methodology. Upon the import of OBO ontologies, ontology files in the OWL format are downloaded from the OBO Foundry website (<http://www.obofoundry.org/>). Then, the ontology is directly imported into SciNetS. Third, in repository projects, the complete data from a database are stored as single or multiple files. As a result, 27 projects (17 for database, nine for ontology and one for repository) composed of 108 396 classes and 777 319 instances were

defined as for September in 2010. These projects are updated monthly in average from constituent databases and ontologies.

Then, we examined the contents and semantics (not the data format or syntax) of 41 classes of imported projects, which play the principal roles in each project. To ensure the consistent classification of the content, we used a top-middle level ontology, YAMATO-GXO Lite (<http://scinets.org/item/rib23i/>), which is the lightened version of the middle-level ontology, Genetics Ontology (GXO) (30) ([http://www.brc.riken.jp/lab/bpmp/ontology/ontology\\_gxo.html](http://www.brc.riken.jp/lab/bpmp/ontology/ontology_gxo.html)), to bridge between the experimental genetics domain and the latest top-level ontology, Yet Another More Advanced Top-level Ontology (YAMATO) (31) ([http://www.ei.sanken.osaka-u.ac.jp/hozo/onto\\_library/upperOnto.htm](http://www.ei.sanken.osaka-u.ac.jp/hozo/onto_library/upperOnto.htm)). YAMATO-GXO Lite was developed with the ontology editor in SciNetS (paper in preparation). As a result, 41 classes conveying the key information from each project are classified under the fifteen upper classes as follows: 'Genome segment and gene in mammal', 'Allele in mammal', 'Transcript in mammal', 'Protein in mammal', 'Strain resource in mammal', 'Cell line resource in mammal', 'Disease', 'Experimental data with mammalian sample' and 'Mammalian Orthologous group' (Figure 3). The RIKEN Integrated Database of Mammals is implemented as a project to define these classes as a root (<http://SciNetS.org/db/mammal>). The ontology-based classification of contents was embodied with `rdf:subclassOf` links, which can be applied across multiple projects in SciNetS. To integrate across species



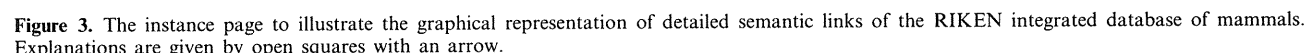
**Figure 2.** The implementation of data in the RIKEN integrated database of mammals to ensure direct integration between ontologies and databases based on the semantic web technology. Individual public and RIKEN databases are imported as individual projects, and their main contents are reviewed to classify them into the lower classes to root the 15 classes of the integrated database, such as gene, transcript, experimental data, strain and so on. The classification follows the top-level ontology and is directly linked to the equivalents of rdfs:subclassOf across projects. Property links also organized with rdfs:subPropertyOf to represent the logical definition of upper classes are inherited to lower classes or instances.

databases, we applied the 'query-class', which dynamically refers only to specific instances from another class. For example, the diffraction data class in the SSBC project includes the diffraction data from mammal and non-mammal proteins. To extract only mammal data, we implemented the query-class, which is an expanded use of the owl:oneOf element to define a class by enumerating its elements. With these operations, the project for the integrated database works as the bridge to connect the YAMATO-GXO Lite and imported projects, in which the imported classes are defined as lower concepts of the top-level ontology as shown in Figure 2.

In the next step, to ensure further semantic integration of the imported data, we examined the equivalencies of property links (semantic links) between the upper ontology and lower classes in imported projects. For example, the 'Allele' class in YAMATO-GXO Lite has a property named 'variant\_of' that takes its value from the range of the 'Genome segment' class. It is the logical representation of one of the features of an allele that the 'allele is a variant of a genome segment'. The

examination of properties in lower classes reveals that the 'MGI allele' class has the 'MGI gene' property range of 'MGI gene', which is equivalent to 'variant\_of'. Consequently, we defined the 'MGI gene' property as a specified type of (rdfs:subPropertyOf) 'variant\_of' to show that *Gdf5<sup>Rgsc451</sup>*, an instance of the MGI allele class, is a variant of *Gdf5*, an instance of the MGI gene class. With this equivalence mapping of properties between YAMATO-GXO Lite and lower imported database classes, we built the ontology-based information structure so that information defined in the upper classes is instantiated in lower database classes and instances.

In addition, regarding the import of external and internal data records, multiple overlaps of records (instances) were collapsed to represent a single identical entity in the real world (i.e. instances of a gene in the Ensembl, MGI and FANTOM projects). We also examined such equality between instances in lower classes that belong to a single upper class. We related identical data items with a semantic link that is equivalent to owl:sameAs.



At the top page of this integrated database, users can overlook all the classes of integrated databases and those data sizes shown in Supplementary Table S2. The overview of the data structure is presented on the ‘data folder’ page, where users can navigate down the class hierarchy across database or ontology projects by clicking on the folder icons that represent classes. On the page of each project, detailed explanations of the projects and URL links to the original database websites are shown. On the class and instance pages, detailed explanations, a table view of instances, a graphic representation of semantic links and links to original data records are displayed (Figure 3).

— 77 —

The RIKEN integrated database of mammals

Links | Wiki | Download | Information | Map

JAPANESE | Log on | Terms of Use

SciNets Gdf5 Search Search All Clear

The RIKEN integrated database of mammals

Home | Concept | Data folders | Search guide | Download guide | Contact

Pos **Attd** Mouse Gene Search Human Gene Search

The followings are filtered by 'Gdf5'

Show all hits (123 hits) Clear

**Disks**

The RIKEN in... (43 hits / 704,204)

RIKEN SciNets... (8 hits / 139,532)

Genetics Ont... (41 hits / 682,425)

Mouse Ensembl Gene (1 hit / 37,077)

Rat Ensembl Gene

MGI Strains an... (5 hits / 11,805)

RIKEN ENR Mouse... (8 hits / 2,023)

Mouse MGI Gene (1 hit / 37,158)

HP Human phenotype ontology

Mammalian Ortholog (1 hit / 36,519)

FMA Foundational Model of Anato...

Mouse pathology

Entrez Mouse (1 hit / 67,344)

RefSeq Mouse (2 hits / 76,574)

Chimpanzee Ense... (1 hit / 24,933)

RefSeq Human (8 hits / 81,422)

HUGO Gene Nomen... (1 hit / 29,385)

Cerebellar Deve... (1 hit / 45,573)

MGI Alleles an... (8 hits / 36,980)

Dog Ensembl Gene (1 hit / 23,550)

Mouse Anatomy Ontology

Human Ensembl Gene (1 hit / 66,234)

Genome Sequence

Rat Genome Data... (1 hit / 39,818)

mammalian phenotype ontology

FANTOM4 (1 hit / 42,851)

QMM (6 hits / 20,025)

Resources in RIKEN Bioresource ...

Enrich Human (1 hit / 42,525)

RAPID: Resource of Primary Immu...

FANTOM eedb (1 hit / 38,971)

**Contents**

Home

Allele in mammal (9 hits / 37,304)

Allele... (9 hits / 37,304)

Allele in human

Genome segmen... (9 hits / 368,004)

Gene L... (3 hits / 138,138)

Gene L... (3 hits / 144,568)

Gene in rat (1 hit / 39,018)

Gene in dog (1 hit / 23,550)

Gene in ... (1 hit / 24,823)

Transcript in ... (5 hits / 84,432)

Transcr... (4 hits / 43,581)

Transcr... (1 hit / 40,871)

Protein in mammal (5 hits / 73,564)

Protein... (4 hits / 37,061)

Protein ... (1 hit / 35,703)

Strain resource... (6 hits / 11,772)


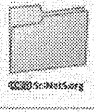
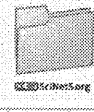
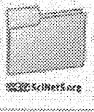




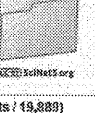

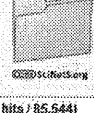


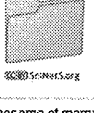
Mouse s... (6 hits / 11,772)

Cell line resource in mammal

Hybrid cell in mammal

Contents of The RIKEN integrated database of mammals

Page(s) 1 1 - 14 of 14

 <b>Home</b> Last update: 2010-09-10	 <b>✓ (9 hits / 37,304)</b> <b>Allele in mammal</b> Last update: 2010-09-11	 <b>✓ (9 hits / 368,004)</b> <b>Genome segment and gene in mammal</b> Last update: 2010-10-14	 <b>✓ (5 hits / 84,432)</b> <b>Transcript in mammal</b> Last update: 2010-07-01
 <b>✓ (5 hits / 73,564)</b> <b>Protein in mammal</b> Last update: 2010-07-01	 <b>✓ (6 hits / 11,772)</b> <b>Strain resource in mammal</b> Last update: 2010-07-02	 <b>Cell line resource in mammal</b> Last update: 2010-07-01	 <b>Phenotype in mammal [ontology]</b> Last update: 2010-07-27
 <b>✓ (6 hits / 18,889)</b> <b>Disease</b> Last update: 2010-07-01	 <b>Anatomical part in mammal [ontology]</b> Last update: 2010-07-01	 <b>✓ (2 hits / 85,544)</b> <b>Experimental data with mammalian sample</b> Last update: 2010-08-26	 <b>✓ (1 hit / 20,815)</b> <b>Mammalian Orthologous group</b> Last update: 2010-07-01
 <b>Clone of mammalian DNA</b> Last update: 2010-09-03	 <b>Chromosome of mammal</b> Last update: 2010-08-26		

Page(s) 1 1 - 14 of 14

**Figure 4.** The representation of filtering (search) result of the RIKEN integrated database of mammals. The number of query hits is represented on each disk or folder icon with red letters.



information, protein–protein interactions, co-expression data, orthologous genes, drugs and metabolite information (32,33). These search functions are implemented by our original database search engine GRASE (28). Data in this database are downloadable from the ‘Download’ links of each project with specifications of licenses via CC or GNU. SciNetS provides various several standard formats, such as RDF, OWL or tab-delimited files.

### MERITS OF THE DIRECT INTEGRATION ONTOLOGY WITH DATABASE

The RIKEN integrated database of mammals should be the first practical database to perform the direct integration of the top-level ontology, domain-specific ontologies and the existing databases. Although there is much room for improvement, this database represents a simple and practical methodology to generate a consistent and scalable body of information that is interoperable with the global informational whole based on semantic web technology. In the process of the integration, we have investigated data schema of each database and classified their contents based on the top-level ontology. These operations are comparable to the ‘annotation’ of databases.

Currently, the main knowledge framework is provided by a top-level ontology, YAMATO-GXO lite. During the development of this ontology, it was optimized to allow the integration of multiple biological databases used by the mammalian genetics community. For example, the basic definition of mammalian genes is provided by the Mouse Genomic Nomenclature Committee (MGNC), which is suitable for data management of genome information. It defines gene as ‘a functional unit, usually encoding a protein or RNA, whose inheritance can be followed experimentally’; also, ‘a gene symbol should be unique within the species’. This definition is surely represented in the MGI database because each gene record is stored in the genome segment (phrased as ‘genetic marker’ in MGI) database as a subset (or a subclass) having a biological function and is unique in the mouse genome. An allele is defined as a variant form of a genome segment, which is usually unique for the sequence of itself. Here, we should mention that there are at least two ways to conceptualize genome segments and alleles. One attaches greater importance to the instantiation toward a molecule. Such a classification may be performed in the BioTop top-level ontology (34). Another applies the conceptualization of gene and allele as classes and allows them to have their own instances such as *Gdf5* and *Gdf5<sup>Rgsc451</sup>*. YAMATO-GXO lite applies latter as useful for integrating databases. A gene is a subclass of the genome segment that has a biological function. An allele is defined as a different class to be unique for conveying information and is equal to the nucleotide sequence.

The consistent knowledge framework contributes to metadata-based and cross-database retrieval for easy and clear specification of the range of the search object. Such retrieval was previously only available for individual databases. For example, to search for ‘the mouse genome segment that has a variant with a point mutation’, a cross-database retrieval is usually performed with the

combination of the text, ‘genome segment’ ‘mouse’ and ‘point mutation’. Such a search never indicates the range of the search resource, ‘genome segment of mouse’, which is a subclass of genome segments of mammals. Furthermore, the range must be clearly distinguished from the mouse allele, which is the entity that has the point mutation. In this database, the fifteen upper classes and the lower class-tree are explicitly defined to represent the range of resources and the organization of metadata. Therefore, the knowledge framework enables the retrieval of specific resources, such as ‘genome segment of mouse’, to be related to the text ‘point mutation’ (which may be described in the instance of an allele) using query languages such as SPARQL or GRASQL. On the GUI of this database, the simple GRASQL-based searches are implemented as simple text searches, as described above.

The knowledge framework also contributes to ensuring the cost-effective sustainability and updating of data. In the implementation of SciNetS, the common body for data integration, the continuous maintenance and management of data are essential. These operations are differentiated with respect to not only the formalism of data but also the contents in each database. The consistently integrated data, which represent classification and inheritances between property links, reveal the content-oriented standardization of the formalism of data items. We are now developing content-oriented procedures for data maintenance specified for data contents such as gene, allele and strain. The standardized data formulation provided from top- and middle- level ontologies reduces the labor cost of data management through the reduction of unevenness in the operations of individual databases. Thus, the ‘annotation’ of databases helps to design the contents-oriented common user interfaces or the procedure of data management of imported databases, which had been independently developed in different research projects.

Another advantage of the data integration on SciNetS is that the continuous improvements and enhancements are ensured by the data tracking system to integrate newly added projects. We are planning to incorporate other mammal-related databases into RIKEN to disseminate them to broad communities. Public data are also incorporated to provide higher usability by establishing relationships among data. For example, we still do not ensure fully functional cross-species integration of anatomies and phenotypes, which are provided as species-specific ontologies. To solve this problem, we need equivalence mapping of homologous organs/tissues and phenotypes. Some ontology developers are working on this issue to establish relationships between the Mammalian Phenotype ontology (MP) (35) and Human Phenotype Ontology (HPO) (36–37) mediated by the Phenotypic Quality Ontology (PATO) (38–44). The implementation of such equivalence information in the integrated database will greatly improve the utility of phenotype data to provide cross-mapping information with diseases. Furthermore, we are also integrating the plant omics data using SciNetS with a similar methodology (K. Doi *et al.* manuscript in preparation).

Referring to the same top-level ontology, we are planning to integrate the mammalian database with the plant one. One of the merits of the institute-oriented data integration is the promotion of data integration across phylogenetically distant species because the species- or community-oriented integration of plant and mammal information is often difficult.

## FUTURE DIRECTIONS

We will continue the development of this database to enhance the data, retrieval functions and semantics as described above. In addition, we are also planning to incorporate other top-middle level ontologies beyond YAMATO-GXO lite, such as the Basic Formal Ontology (BFO) (45), the Descriptive Ontology for Linguistic, Cognitive Engineering (DOLCE) (46), BioTop and the Ontology of Biomedical Investigation (OBI). In YAMATO, the interoperability among these top-level ontologies represents a general model to explain differentiation and interrelationships among classes (31). With this enhancement, we will cooperate with the global efforts of the OBO Foundry, the initiative activity of the OBO consortium, which has been to coordinate the scientific methods in ontology developments toward forming a consistent, cumulatively expanding and algorithmically tractable whole (7) based on the BFO as the semantic framework.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

The authors thank Drs Kaoru Saijyo, Kazuyuki Mekada and Hatsumi Nakata in RIKEN BRC to help data import from Resource database to SciNetS.

## FUNDING

Maintenance of SciNetS is supported by the Integrated Database Project by Ministry of Education, Culture, Sports, Science and Technology (MEXT).

*Conflict of interest statement.* None declared.

## REFERENCES

- Abbott, A. (2009) Plant genetics database at risk as funds run dry. *Nature*, **462**, 258–259.
- Maltais, L.J., Blake, J.A., Eppig, J.T. and Davisson, M.T. (1997) Rules and guidelines for mouse gene nomenclature: a condensed version. International Committee on Standardized Genetic Nomenclature for Mice. *Genomics*, **45**, 471–476.
- Wain, H.M., Lush, M., Ducluzeau, F. and Povey, S. (2002) Genew: the human gene nomenclature database. *Nucleic Acids Res.*, **30**, 169–171.
- Twigger, S.N., Shimoyama, M., Bromberg, S., Kwik, A.E. and Jacob, H.J. (2007) RGD Team. The rat genome database, update 2007—easing the path from disease to data and back again. *Nucleic Acids Res.*, **35**, D658–D662.
- 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.
- 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.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S. and Eppig, J.T. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
- Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.
- Mouse Phenotype Database Integration Consortium, Hancock, J.M., Adams, N.C., Aidinis, V., Blake, A., Bogue, M., Brown, S.D., Chesler, E.J., Davidson, D., Duran, C. *et al.* (2007) Mouse Phenotype Database Integration Consortium: integration of mouse phenotype data resources. *Mamm. Genome*, **18**, 157–163.
- Chandras, C., Weaver, T., Zouberakis, M., Smedley, D., Schughart, K., Rosenthal, N., Hancock, J.M., Kollias, G., Schofield, P.N. and Aidinis, V. (2009) Models for financial sustainability of biological databases and resources. *Database*, doi:10.1093/database/bap017.
- Schofield, P.N., Bubela, T., Weaver, T., Portilla, L., Brown, S.D., Hancock, J.M., Einhorn, D., Tocchini-Valentini, G., Hrabe de Angelis, M., Rosenthal, N. and CASIMIR Rome Meeting participants. (2009) Post-publication sharing of data and tools. *Nature*, **461**, 171–173.
- Schatz, M.C., Langmead, B. and Salzberg, S.L. (2010) Cloud computing and the DNA data race. *Nat. Biotechnol.*, **28**, 691–693.
- Berners-Lee, T., Hendler, J. and Lassila, O. (2001) The semantic web. *Scientific American*, May, pp. 29–37.
- FANTOM Consortium, Suzuki, H., Forrest, A.R., van Nimwegen, E., Daub, C.O., Balwierz, P.J., Irvine, K.M., Lassmann, T., Ravasi, T., Hasegawa, Y. *et al.* (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.*, **41**, 553–562.
- Kawaji, H., Severin, J., Lizio, M., Forrest, R.R.A., Nimwegen, E., Rehli, M., Shroder, K., Irvine, K., Suzuki, H., Carninci, P. *et al.* (2011) Update of FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Res.* (in press).
- Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.
- Kagami, Y. and Furuichi, T. (2001) Investigation of differentially expressed genes during the development of mouse cerebellum. *Brain Res. Gene Expr. Patterns*, **1**, 39–59.
- Sato, A., Sekine, Y., Saruta, C., Nishibe, H., Morita, N., Sato, Y., Sadakata, T., Shinoda, Y., Kojima, T. and Furuichi, T. (2008) Cerebellar development transcriptome (CDT-DB): profiling of spatio-temporal gene expression during the postnatal development of mouse cerebellum. *Neural Networks*, **21**, 1056–1069.
- Yoshiki, A., Ike, F., Mekada, K., Kitaura, Y., Nakata, H., Hiraiwa, N., Mochida, K., Ijuin, M., Kadotani, M., Murakami, A. *et al.* (2009) The mouse resources at the RIKEN BioResource center. *Exp. Anim.*, **58**, 85–96.
- Nakamura, Y. (2010) Bio-resource of human and animal-derived cell materials. *Exp. Anim.*, **59**, 1–7.
- Yokoyama, K.K., Murata, T., Pan, J., Nakade, K., Kishikawa, S., Ugai, H., Kimura, M., Kujime, Y., Hirose, M., Masuzaki, S. *et al.* (2010) Genetic materials at the gene engineering division, RIKEN BioResource Center. *Exp. Anim.*, **59**, 115–124.
- Masuya, H., Nakai, Y., Motegi, H., Niinaya, N., Kida, Y., Kaneko, Y., Aritake, H., Suzuki, N., Ishii, J., Koorikawa, K. *et al.* (2004) Development and implementation of a database system to manage a large-scale mouse ENU-mutagenesis program. *Mamm. Genome*, **15**, 404–411.



23. Masuya,H., Yoshikawa,S., Heida,N., Toyoda,T., Wakana,S. and Shiroishi,T. (2007) Phenosite: a web database integrating the mouse phenotyping platform and the experimental procedures in mice. *J. Bioinform. Comput. Biol.*, **5**, 1173–1191.
24. Keerthikumar,S., Raju,R., Kandasamy,K., Hijikata,A., Ramabadran,S., Balakrishnan,L., Ahmed,M., Rani,S., Selvan,L.D., Somanathan,D.S. *et al.* (2009) RAPID: Resource of Asian Primary Immunodeficiency Diseases. *Nucleic Acids Res.*, **37**, D863–D867.
25. Hijikata,A., Kitamura,H., Kimura,Y., Yokoyama,R., Aiba,Y., Bao,Y., Fujita,S., Hase,K., Hori,S., Ishii,Y. *et al.* (2007) Construction of an open-access database that integrates cross-reference information from the transcriptome and proteome of immune cells. *Bioinformatics*, **23**, 2934–2941.
26. Bono,H., Kasukawa,T., Hayashizaki,Y. and Okazaki,Y. (2002) READ: RIKEN Expression Array Database. *Nucleic Acids Res.*, **30**, 211–213.
27. Eppig,J.T. and Strivens,M. (1999) Finding a mouse: the International Mouse Strain Resource (IMSR). *Trends Genet.*, **15**, 81–82.
28. Rubin,D.L., Noy,N.F. and Musen,M.A. (2007) Protégé: a tool for managing and using terminology in radiology applications. *J. Digit. Imaging.*, **20**, 34–46.
29. Kobayashi,N. and Toyoda,T. (2008) Statistical search on the Semantic Web. *Bioinformatics*, **24**, 1002–1010.
30. Masuya,H. and Mizoguchi,R. (2009) Toward fully integration of mouse phenotype information. *Proceedings of the Second Interdisciplinary Ontology Meeting*, Keio University Press, February 28–March 1, 2009, Tokyo, Japan, pp. 35–44.
31. Mizoguchi,R. (2009) Yet Another Top-level Ontology: YATO. *Proceedings of the Second Interdisciplinary Ontology Meeting*, Keio University Press, February 28 - March 1, 2009, Tokyo, Japan, pp. 91–101.
32. Yoshida,Y., Makita,Y., Heida,N., Asano,S., Matsushima,A., Ishii,M., Mochizuki,Y., Masuya,H., Wakana,S., Kobayashi,N. *et al.* (2009) PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Res.*, **37**, W147–W152.
33. Makita,Y., Kobayashi,N., Mochizuki,Y., Yoshida,Y., Asano,S., Heida,N., Deshpande,M., Bhatia,R., Matsushima,A., Ishii,M. *et al.* (2009) PosMed-plus: an intelligent search engine that inferentially integrates cross-species information resources for molecular breeding of plants. *Plant Cell Physiol.*, **50**, 1249–1259.
34. Schulz,S., Beisswanger,E., van den Hoek,L., Bodenreider,O. and van Mulligen,E.M. (2009) Alignment of the UMLS semantic network with BioTop: methodology and assessment. *Bioinformatics*, **25**, i69–i76.
35. Smith,C.L., Goldsmith,C.A. and Eppig,J.T. (2005) The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.*, **6**, R7.
36. Robinson,P.N. and Mundlos,S. (2010) The human phenotype ontology *Clin. Genet.*, **77**, 525–534.
37. Köhler,S., Schulz,M.H., Krawitz,P., Bauer,S., Dölken,S., Ott,C.E., Mundlos,C., Horn,D., Mundlos,S. and Robinson,P.N. (2009) Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.*, **85**, 457–464.
38. Gkoutos,G.V., Green,E.C., Mallon,A.M., Blake,A., Greenaway,S., Hancock,J.M. and Davidson,D. (2004) Ontologies for the description of mouse phenotypes. *Comp. Funct. Genomics.*, **5**, 545–551.
39. Gkoutos,G.V., Green,E.C., Mallon,A.M., Hancock,J.M. and Davidson,D. (2005) Using ontologies to describe mouse phenotypes. *Genome Biol.*, **6**, R8.
40. Washington,N.L., Haendel,M.A., Mungall,C.J., Ashburner,M., Westerfield,M. and Lewis,S.E. (2009) Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol.*, **7**, e1000247.
41. Beck,T., Morgan,H., Blake,A., Wells,S., Hancock,J.M. and Mallon,A.M. (2009) Practical application of ontologies to annotate and analyse large scale raw mouse phenotype data. *BMC Bioinformatics*, **6**(Suppl. 5), S2.
42. Mungall,C.J., Gkoutos,G.V., Smith,C.L., Haendel,M.A., Lewis,S.E. and Ashburner,M. (2010) Integrating phenotype ontologies across multiple species. *Genome Biol.*, **11**, R2.
43. Schofield,P.N., Gkoutos,G.V., Gruenberger,M., Sundberg,J.P. and Hancock,J.M. (2010) Phenotype ontologies for mouse and man: bridging the semantic gap. *Dis. Model Mech.*, **3**, 281–289.
44. Hancock,J.M., Mallon,A.M., Beck,T., Gkoutos,G.V., Mungall,C. and Schofield,P.N. (2010) Mouse, man, and meaning: bridging the semantics of mouse phenotype and human disease. *Mamm. Genome*, **20**, 457–461.
45. Grenon,P. and Smith,B. (2004) SNAP and SPAN: towards dynamic spatial ontology. *Spat. Cogn. Comput.*, **4**, 69–103.
46. Gangemi,A., Guarino,N., Masolo,C., Oltramari,A. and Schneider,L. (2002) Sweetening ontologies with DOLCE, knowledge engineering and knowledge management. *Lecture Notes In Computer Science Vol. 2473, 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*. Springer, London, UK, pp. 166–181.

The Sonoda-Tajima Cell Collection, a human genetics research resource with emphasis  
on South American indigenous populations

Inaho Danjoh<sup>1</sup>, Kaoru Saijo<sup>1</sup>, Takashi Hiroyama<sup>1</sup>, and Yukio Nakamura<sup>1,\*</sup>

<sup>1</sup>Cell Engineering Division, RIKEN BioResource Center

Koyadai 3-1-1, Tsukuba, Ibaraki, 305-0074, Japan

\*Author for Correspondence: Yukio Nakamura, Cell Engineering Division, RIKEN  
BioResource Center

TEL +81 29 836 9124

FAX +81 29 836 9049

E-mail; yukionak@brc.riken.jp

## ABSTRACT

The Sonoda-Tajima Cell Collection includes cell samples obtained from a range of ethnic minority groups across the world but in particular from South America. The collection is made all the more valuable by the fact that some of these ethnic populations have since died out, and thus it will be impossible to prepare a similar cell collection again. The collection was donated to our institute, a public cell bank in Japan, by Drs. Sonoda and Tajima to make it available to researchers throughout the world. The original cell collection was composed of cryopreserved peripheral blood samples that would obviously have been rapidly exhausted if used directly. We, therefore, immortalized some samples with the Epstein-Barr Virus and established B lymphoblastoid cell lines (B-LCLs). As there is continuing controversy over whether the B-LCL genome is stably maintained, we performed an array comparative genomic hybridization (CGH) analysis to confirm the genomic stability of the cell lines. The array CGH analysis of the B-LCL lines and their parental B cells demonstrated that genomic stability was maintained in the long-term cell cultures. The B-LCLs of the Sonoda-Tajima Collection will therefore be made available to interested scientists around the world. At present, 512 B-LCLs have been developed, and we are willing to increase the number if there is sufficient demand.

## KEYWORDS

Amerind, minority group, B-LCL, array CGH

## INTRODUCTION

Human T-lymphotropic virus type 1 (HTLV-I) is the causative virus of adult T cell leukemia (ATL) (Uchiyama et al., 1977) and HTLV-I-associated chronic myelopathy (HAM) (Osame et al., 1986). HTLV-I is distributed worldwide and is phylogenetically classified into three major subtypes: the Central African group detected in the central African continent; the Melanesian group located around Australia; and, the Cosmopolitan (Mongoloid) group widely spread across the Asian region of the Eurasian continent (reviewed by Yamashita et al., 1996, and by Sonoda et al., 2011). HTLV-II, which is closely related to HTLV-I but is a distinct virus, is also classified into two major groups, HTLV-IIa and HTLV-IIb. HTLV-II is also detected worldwide and its distribution shows some geographic bias, i.e., HTLV-IIa0 and a3 were predominantly detected in non-indigenous populations in North America; a4 and b1 were specifically detected in indigenous people in North and South America, respectively; b5 was mainly detected in indigenous populations in both North and the South America; only a2 and b4 were detected within European populations (Switzer et al., 1995).

To confirm the reported patterns of geographic and ethnic segregation of human HTLV-I and -II in human populations, Sonoda, Tajima and colleagues conducted seroepidemiological studies of indigenous populations in South America who lived in closed societies (Komurian-Pradel et al., 1992; Ijichi et al., 1993; Miura et al., 1994; Miura et al., 1997; Li et al., 1999; Fujiyoshi et al., 1999). In this series of analyses that included both extant populations and the preserved remains of prehistoric mummies, they showed that the Amerind populations retained the Mongoloid subtype of HTLV-I (Miura et al., 1994; Miura et al., 1997; Li et al., 1999; reviewed by Sonoda et al., 2011) and a distinctive subclass of HTLV-IIb (Ichiji et al. 1993; Miura et al., 1997). They also showed that there was a geographic bias in the distribution of HTLV-I/II carriers: HTLV-I predominated in the Andes highlands, while there were foci of HTLV-II in the lowlands of South America. From these results, they concluded that ancestors of the Amerind populations carried HTLV-I and -II into the South American continent from the Eurasian continent over 10,000 years ago and that the indigenous South American populations could be divided into two major ethnic groups.

During the course of these studies, a number of peripheral blood samples were obtained and cryopreserved with the consent of the donors, not only for immediate use but also for future studies. In addition to these samples, many other peripheral blood

samples were collected from isolated ethnic populations in various areas around the world (Figure 1). All of these samples have been donated to a not-for-profit public cell bank held at the Cell Engineering Division of RIKEN BioResource Center in Tsukuba, Japan. Overall, more than 3,500 blood samples were donated to the cell bank.

One obvious problem with these cryopreserved peripheral blood samples is that if they are used for experimental studies, then they would quickly run out. Since these samples are an extremely precious resource for future research in human genetics, it was decided that they should be preserved in a form that could be expanded repeatedly. One well-established method is to transform B lymphoid cells in peripheral blood using the Epstein-Barr virus (EBV) (Nilsson, 1979). The genomic stability of B lymphoblastoid cell lines transformed by EBV (B-LCLs) has been evaluated, such as by karyotyping using conventional G-band staining (reviewed by Nilsson, 1992; Okubo et al., 2001), and by analysis of particular genetic loci for mutations (Lalle et al., 1995). Those analyses indicate that the genome of these B-LCLs is stable. Recent technical advances have now opened up the possibility of a more stringent evaluation of genomic stability in B-LCLs. For example, Simon-Sanchez et al. (2006) and Herbeck et al. (2009) compared genome-wide single nucleotide polymorphism (SNP) patterns in B-LCLs and parental B cells (i.e., the original cells that B-LCLs were derived from), and concluded



that there were no statistically significant differences between the cell types. By contrast, the Wellcome Trust Case Control Consortium analyzed copy number variation (CNV) over 3,400 loci and detected differences between B-LCLs and their parental cells at a significant number of loci (The Wellcome Trust Case Control Consortium, 2010).

To analyze the genomic stability of B-LCLs, we performed array CGH (comparative genomic hybridization) analysis using eleven B-LCLs and their parental cells. This analysis confirmed the stability of the genomes of the B-LCLs. We have, therefore, now established more than 500 B-LCLs from the cryopreserved cell samples of the Sonoda-Tajima Collection. The samples used for establishing B-LCLs were selected to include as many ethnic populations from South America as possible.

## **MATERIALS AND METHODS**

### **Peripheral blood samples**

The ethnic populations who kindly donated peripheral blood samples following informed consent and the numbers of individuals involved are given in Table 1. The approximate geographic locations where the samples were collected are shown in Figure 2. Peripheral blood mononuclear cells (PBMNCs) were separated from each blood sample and cryopreserved in liquid nitrogen until use in this study. The ethical

committee of the RIKEN Tsukuba Institute approved the use of these samples before the study was initiated.

### **Establishment of B-LCLs**

B-LCLs were established using previously reported methods (Bird et al., 1981; Rickinson et al., 1984). The B95-8 cell line was obtained from the Cell Resource Center for Biomedical Research, Tohoku University (Sendai, Miyagi, Japan) and cultured in RPMI1640 (Gibco, Carlsbad, CA, USA) supplemented with 10% fetal bovine serum (FBS). The culture supernatant of the B95-8 cells was collected, filtered to remove cells, cryopreserved, and used as the source of EBV. For infection of PBMNCs with EBV, the B95-8 culture supernatant was thawed and incubated with the PBMNCs at 37°C for 2 hours. The cells were then washed with RPMI, resuspended in RPMI1640 supplemented with 20% FBS and 0.5 µg/ml cyclosporin A (trade name Sandimmun; Novartis Pharma, Basel, Switzerland), inoculated into a multi-well plate at a cell density of approximately  $2 \times 10^5$  cells/cm<sup>2</sup>, and cultured. Half of the medium was changed twice a week with replacement by fresh medium. After a few weeks and upon confirming efficient proliferation of the cells, the cultures were scaled up using a 2- to 4-fold dilution into 75 cm<sup>2</sup> culture flasks. B-LCLs around passage 10 were deposited in the

cell bank of the Cell Engineering Division of RIKEN BioResource Center (RIKEN Cell Bank) and were used for the following analyses.

### **Microsatellite polymorphism analysis**

To authenticate the identity of each cell line, short tandem repeat (STR) polymorphisms in microsatellites were analyzed in genomic DNA using the PowerPlex1.2 kit (Promega, Madison, WI, USA). This PCR-based analysis kit includes the primer sets required to detect STR polymorphisms at eight loci (Masters et al., 2001; Yoshino et al., 2006).

### **Karyotype analysis**

Chromosome preparations were made in a standard fashion and then G-banded (Yunis et al., 1978). Chromosome numbers were counted in 50 cells (mode-analysis), and then the G-band pattern was analyzed in detail in 20 of the cells to identify chromosome aberrations (karyotype analysis). These analyses were performed for us by Nihon Gene Research Laboratories (Sendai, Miyagi, Japan).

### **Collection of B-lineage cells from blood samples**

The anti-FITC MultiSort Kit (Miltenyi Biotech, Bergisch Gladbach, Germany) and a

MACS MS column (Miltenyi Biotech) were used to collect CD19-positive (CD19<sup>+</sup>) B-lineage cells from PBMNCs and umbilical cord blood mononuclear cells (CBMNCs) according to the manufacturer's instructions with slight modification. Briefly, to remove any dead cells, approximately  $5 \times 10^7$  cells were passed through the column without staining with antibodies. Phosphate buffered saline (PBS) supplemented with 0.5% FBS and 0.05% sodium azide was used to wash the column. The collected viable cells were stained with FITC (fluorescein isothiocyanate)-labeled anti-human CD19 antibody (BD Biosciences, San Jose, CA, USA) and then reacted with anti-FITC MultiSort beads. We collected CD19<sup>+</sup> B-lineage cells attached to MultiSort beads. After removal of the beads by proteolytic cleavage, genomic DNA was extracted from the cells. Cell numbers at each step were counted with a hemacytometer, and the purity of the CD19<sup>+</sup> cells was analyzed with a FACS Calibur flow cytometer (BD Biosciences). On average, approximately  $5 \times 10^5$  CD19<sup>+</sup> cells were collected from  $5 \times 10^7$  PBMNCs.

#### **Array CGH analysis**

Genomic DNA was obtained from the cells using a DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany). Preparation of the fluorescent-labeled samples for array CGH analysis was performed according to the manufacturer's instructions. Briefly, 1.0 -

1.5 µg of genomic DNA was digested with the restriction endonucleases AluI and RsaI. Then, genomic DNAs from PBMNCs and CD19-negative (CD19<sup>-</sup>) cells were labeled with cyanine 3-dUTP (Cy3) and genomic DNAs from B-LCLs and CD19<sup>+</sup> cells were labeled with cyanine 5-dUTP(Cy5) using a Genomic DNA Enzymatic Labeling Kit (Agilent, Santa Clara, CA, USA). After evaluating labeling efficiency, Cy3-labeled and Cy5-labeled DNAs were mixed in the Oligo aCGH/ChIP-on chip Hybridization Kit (Agilent) and then hybridized on a Sure Print G3 Human CGH 2x400K microarray (Agilent) at 65°C for 40 hours in a hybridization oven, with rotation at 20 rpm. The microarrays were scanned with a DNA Microarray Scanner (Agilent) at 3 µm resolution. To quantify the intensity of the fluorescent signal of each spot on the microarray, we used Feature Extraction software version 10.5.1.1 (Agilent).

### **Statistical analysis**

The array CGH analysis data were evaluated to identify genomic alterations by a statistical analysis with Genomic Workbench Standard Edition version 5.0.14 (Agilent). The conditions and parameters of the statistical analysis were as follows: the Moving Average (Log ratio) algorithm was linear at a 2 Mb window size, the ADM-2 algorithm threshold was set to 6.0 with Fuzzy Zero, the aberration filters were applied at a

minimum number of probes of 3 in the region and 0.5 minimum absolute average log ratio for the region. After the first screening described above, the raw data of selected loci were individually scrutinized. As a result, some "aberrations" were removed from the aberration list because they were located within a noisy area, such as the telomeres, and the reliability of the aberration calls in such region was low. Although Genomic Workbench Standard Edition version 5.0.14 adopts hg18 for gene mapping, we referred to Build36.3 in the Map Viewer constructed by NCBI (<http://www.ncbi.nlm.nih.gov/projects/mapview/>) for detailed mapping of genes.

#### **V(D)J recombination analysis**

Recombination in the V(D)J region of the immunoglobulin (Ig) heavy chain gene in B-LCLs was analyzed by PCR as described previously with slight modification (Kiyoi et al., 1992; Abe et al., 1994). The primers used for the PCR were 5'-GAG TCG AC(A/T) C(A/G)G C(G/CXG/A)T GTA (T/C)T(T/A) CTG-3' for the V common region and 5'-CCA AGC TTA CCT GAG GAG ACG GTG A-3' for the J common region. PCR was performed using an ExTaq polymerase kit (TaKaRa Bio, Otsu, Shiga, Japan) with a reaction mixture containing 0.4  $\mu$ M of the J common primer and 4  $\mu$ M of the V common primer. An initial incubation at 94°C for 5 min was followed by 40 cycles of