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 $Gdf5^{Rgsc451}$. 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

1. Abbott,A. (2009) Plant genetics database at risk as funds run dry. *Nature*, **462**, 258–259.
2. 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.
3. Wain,H.M., Lush,M., Ducluzeau,F. and Povey,S. (2002) Genew: the human gene nomenclature database. *Nucleic Acids Res.*, **30**, 169–171.
4. Twigger,S.N., Shimoyama,M., Bromberg,S., Kwitek,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.
5. 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.
6. 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.
7. 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.
8. 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.
9. Mouse Phenotype Database Integration Consortium, Hancock,J.M., Adams,N.C., Aidinis,V., Blake,A., Bogue,M., Brown,S.D., Chesterm,E.J., Davidson,D., Duran,C. *et al.* (2007) Mouse Phenotype Database Integration Consortium: integration of mouse phenome data resources. *Mamm. Genome.*, **18**, 157–163.
10. 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.
11. 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.
12. Schatz,M.C., Langmead,B. and Salzberg,S.L. (2010) Cloud computing and the DNA data race. *Nat. Biotechnol.*, **28**, 691–693.
13. Berners-Lee,T., Hendler,J. and Lassila,O. (2001) The semantic web. *Scientific American*, **May**, pp. 29–37.
14. 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.
15. Kawaji,H., Severin,J., Lizio,M., Forrest,R.R.A., Nimwegen,vE., Rehli,M., Shroder,K., Irvine,K., Susuki,H., Carninci,P. *et al.* (2011) Update of FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Res.* (in press).
16. 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.
17. 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.
18. 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.
19. Yoshiki,A., Ike,F., Mekada,K., Kitaura,Y., Nakata,H., Hiraiwa,N., Mochida,K., Ijuin,M., Kadotam,M., Murakami,A. *et al.* (2009) The mouse resources at the RIKEN BioResource center. *Exp. Anim.*, **58**, 85–96.
20. Nakamura,Y. (2010) Bio-resource of human and animal-derived cell materials. *Exp. Anim.*, **59**, 1–7.
21. 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.
22. 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. HancockJ.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.

Genome Biology and Evolution Advance Access published March 7, 2011 doi:10.1093/gbe/evr014

Genome Biology and Evolution

SMBE

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

Inaho Danjoh[1], Kaoru Saijo[1], Takashi Hiroyama[1], and Yukio Nakamura[1,*]

[1]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 $2x10^5$ cells/cm^2, 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^2 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 ($CD19^+$) 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 $5x10^7$ 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 $CD19^+$ 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 hematocytometer, and the purity of the $CD19^+$ cells was analyzed with a FACS Calibur flow cytometer (BD Biosciences). On average, approximately $5x10^5$ $CD19^+$ cells were collected from $5x10^7$ 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 ($CD19^-$) cells were labeled with cyanine 3-dUTP (Cy3) and genomic DNAs from B-LCLs and $CD19^+$ 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 at 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 μM of the J common primer and 4 μM of the V common primer. An initial incubation at 94°C for 5 min was followed by 40 cycles of

denaturation at 94°C for 10 sec, annealing at 60°C for 1 sec, and extension at 72°C for 15 sec, and finally incubation at 72°C for 5 min. The PCR products were separated on 3% NuSieve GTG agarose gels (Lonza, Basel, Switzerland) and visualized by ethidium bromide staining. Expected product sizes from recombined V(D)J region were in the range 50-200 bp.

RESULTS

The Sonoda-Tajima Cell Collection

The Sonoda-Tajima cell collection contains more than 3,500 PBMNC samples obtained from ethnic populations across the world (Figure 1). The collection is particularly rich in samples from South America. As stated earlier, any direct use of the PBMNC samples would soon lead to their exhaustion. In comparison to the in vitro expansion of the whole genome using PCR, establishing cell lines is a much better method for maintaining the whole genome in a stable manner. With regard to establishing cell lines from PBMNCs, the generation of B cell lines using EBV (B-LCLs) is the most common approach. The genomic DNA of B-LCLs is maintained in a stable fashion. These facts encouraged us to establish B-LCLs using the PBMNCs of the Sonoda-Tajima Cell Collection. In order to include as many ethnic groups as possible from South America

and a sufficient number of individuals in each group, we selected the samples indicated in Table 1 and Figure 2.

When we used either fresh or cryopreserved PBMNCs in good condition to establish the B-LCLs, we were successful in all cases. However, the PBMNC samples in the Sonoda-Tajima Cell Collection have been cryopreserved for over 20 years. In addition, in sampling areas where no electricity was available, the samples were kept at relatively high temperatures prior to cryopreservation. As a consequence of these factors the viability of some of the thawed PBMNC samples was low, i.e., the cells were in poor condition after thawing, and thus it was not easy to establish B-LCLs for these samples. Overall, a success rate of approximately 80% was achieved for the establishment of B-LCLs from the PBMNC samples of the Sonoda-Tajima Cell Collection. Since the success rate was less than 100%, we decided not to use samples for which there were only one or two tubes cryopreserved. Future treatment of such samples remains to be considered. At the moment, we have established 512 B-LCLs from the Sonoda-Tajima Cell Collection. Information on the established B-LCLs is given in Table 1 and Figure 2.

Karyotype analysis

In contrast to other cell lines, such as immortalized cancer cell lines, one of the most significant features of B-LCLs is their chromosome stability. B-LCL cells retain a normal chromosome karyotype even after relatively long-term culture (reviewed by Nilsson, 1992; Lalle et al., 1995; Okubo et al., 2001); thus, they have been widely used for genetic analysis in many research fields. To confirm that the B-LCLs established from the Sonoda-Tajima Cell Collection maintained a normal karyotype, three B-LCL cell lines, WY084, YAN3191 and YAN3268, were karyotyped following G-banding of chromosome preparations. Chromosome numbers were counted in 50 cells (mode-analysis), and then a detailed G-band analysis was performed in 20 of these cells to identify chromosome aberrations. Consistent with previous reports, the vast majority of the cells had a normal karyotype (Figure 3). In the YAN3268 cell line, all of the cells analyzed had a normal chromosome number as well as a normal karyotype (Figure 3A). In the WY084 cell line, 2 of the 50 cells had a reduced chromosome number (Figure 3B). This might have been an artifact of preparation as no structural aberrations were observed in the cells. In the YAN3191 cell line, all 20 karyotyped cells had an elongated pericentromeric region on the long arm of chromosome 1 (Figure 3C). The centric heterochromatin of chromosome 1 is known to be variable and show heteromorphism between individuals (http://www.rerf.or.jp/dept/genetics/giemsa_5_e.html).

In general, if an irregular karyotype is detected, it is not possible to conclude that it arose during culture because there is a possibility it was already present in the cells of the individual from whom the initial sample was obtained, as in the YAN3191 cell line. In addition, since karyotype analysis has a limited resolution (3-5 Mbp) we cannot detect smaller rearrangements. These limitations prompted us to perform the array CGH analysis on the B-LCLs and their parental PBMNCs.

Array CGH analysis

To avoid the influence of genetic background, such as copy number variation (CNV), we decided to compare the B-LCL and parental PBMNC from the same person in our experiments. We analyzed ten B-LCLs derived from the Sonoda-Tajima Cell Collection and one B-LCL derived from a healthy Japanese volunteer.

The probe set used in the microarray was uninformative regarding the heterochromatic region of chromosome 1. However, in the YAN3191 cell line we did not detect any variants with respect to the other chromosome 1 probes suggesting that there was no major rearrangement of the chromosome (or of the other chromosomes) and that the cell line essentially retained the innate genomic structure (Figure 4).

A few aberrations were observed in all eleven samples: a deletion in

chromosome 14 (in the variable, diversity or joining regions of the Ig heavy chain); a deletion in chromosome 2 (in the variable or joining regions of the Ig κ light chain); an amplification in chromosome 14 (in the variable or joining regions of the T cell receptor (TCR) α chain); and, an amplification in chromosome 7 (in the variable, joining or constant regions of the TCR γ chain). Typical examples of chromosome 14 are shown in Figure 5. In addition, a few other aberrations were detected in some but not all cell lines: a deletion at the immunoglobulin λ locus on chromosome 22; and, amplification at the TCR β on chromosome 7. A precise description of each aberration locus and aberration type described above is given in Table 2 and Figure 6. In addition, a few other specific aberrations were observed in some but not all B-LCLs (Figure S1-S4).

In the B-LCLs, it is possible that the deletions detected in the regions associated with the Ig genes were already present in the B lymphocytes prior to EBV infection. Similarly, since the population of B lineage cells in PBMNCs is very small, it is possible that the detected aberrations were already present in the parent B lymphocytes but not in the non-B cells. To examine this latter possibility, we attempted to compare the genomes of B lineage and non-B lineage cells. However, it was impossible to obtain sufficient B lineage cells from the Sonoda-Tajima Collection. Thus, we used a blood sample from one healthy Japanese adult volunteer and two umbilical

cord blood samples obtained from two independent Japanese neonates. $CD19^+$ and $CD19^-$ cell populations were collected using the magnetic beads system described earlier (see Materials and methods). After selection, the proportion of $CD19^+$ cells was approximately 90% (Figure 7B, C).

As expected, most aberrations observed in B-LCL cell lines were also observed in all $CD19^+$ cells: deletions at the Ig heavy chain on chromosome 14 and at the Ig κ and λ light chains on chromosomes 2 and 22; amplifications at the TCR α locus on chromosome 14, and at the β and γ loci on chromosome 7. A precise description of each aberration locus and aberration type is also described in Table 2 and Figure 6.

Since a flow cytometry analysis indicated that the major cell population in the PBMNCs was $CD3^+$ T-lineage cells (Figure 7A), the results of the statistical analysis of array CGH that suggested "amplification" at the T cell receptor loci were highly likely due to deletions at these loci in T-lineage cells rather than amplification in B-lineage cells.

V(D)J recombination analysis

The array CGH analysis clearly demonstrated that B-LCLs possessed rearranged Ig genes. To confirm the rearrangement of the Ig heavy chain, we performed a PCR