in the contents of the databases and in the services offered are being introduced.

Here, we outline some details of the BioResourceWorld (BRW) integrated database and several representative component databases. All the information described is available at http://www.nbrp.jp/.

NBRP Information Site is a gateway to access all the information of NBRP. The menu on the left-hand side beneath the heading 'Resource Center' provides links to all 30 individual databases, and gives the user direct access to the appropriate database for the organism of interest. The text box 'Keyword' near the upper center of the page is a query box for entering keyword searches of the BRW integrated database. Users can access the home page of BRW directly by clicking on 'Resource Integrated Search Site' in the menu bar near the top of the page. This menu bar also provides links to other pages: 'Journal' provides a link to the Research Resource Circulation (RRC) Web site for browsing and submission of papers reporting research using resources obtained through NBRP; 'Japan Genetic Resources' provides a link to a list of URLs for resources in Japan, but outside the NBRP group; and 'Worldwide Genetic Resources' provides a link to a list of URLs for bioresources worldwide. The total number of records currently available in the database for all bioresources and the subtotals for three individual classifications (animals, plants and microbes) are shown in the turquoise-colored box near the top left of the page.

NBRP DATABASES

An integrated NBRP database retrieval system: BioResourceWorld

BRW is an integrated database retrieval system that allows users to retrieve resources by using the body of information held by NBRP on a number of organisms. All the resources (~4.5 million items) are available for distribution. Keyword searching of BRW is directly available from the NBRP home page. A user performing a search can specify an organism and a resource category from pull-down menus. (The default values are 'All organisms' and 'All categories'.) When a user enters keywords and performs a search with the default settings, the total number of results is shown together with a breakdown of the number of strains and the number of DNA clones. This is followed by the list of resources broken down by organism. By clicking on a desired resource on the list, the user can access a detailed information page that provides links to the distribution-request site for each database (where this exists).

The BRW home page, which is linked from the top menu bar of the NBRP home page, contains the search functions described above and four tabs (ALL, DNA, BLAST and Gene Ontology). These four tabs provide, respectively, a summary display of all resources by organism groups, a summary display of DNA clones by organism groups, a BLAST search and a Gene Ontology (GO) search.

The BLAST search allows the user to search all organism groups and also to specify particular organisms or categories, such as genomic clones, cDNA clones, or libraries. The search results are shown as diagrams of alignments in which abbreviations of names of organisms or groups of organisms appear in a color-coded manner on the left-hand side of the diagram to help users recognize particular organisms. Individual sequences in hits have links to NCBI (1), DDBJ (2) and BRW, which gives the user access to the web sites from which the appropriate resources can be ordered.

The latest service of BRW is a search of resources on the basis of GO. The user can examine the hierarchical structure of a GO term and query it by using GO ID, GO term, or GO Gene. For example, if a user enters '0000038' (or 'very-long-chain fatty acid metabolic process') into the GO-ID (or GO Term), a resource list containing 'mouse (5/5), drosophila (2/4), arabidopsis (10/10)' is shown. The numbers in parenthesis indicate the number of resources associated with this GO-ID (left) and its descendants (right). Thus this service allows users a more semantic search and provides them with a wider range of resources.

The GO search is currently at a testing stage, because not all resources contain GO information. The total number of resources mapped to a GO Term is displayed, and a list of relevant resources can be obtained by clicking the figures. Like other search modes, the GO allows the ordering of resources from search results. We plan to enhance the search of resources across organism by the addition of supporting bio-ontology data, such as phenotype, anatomy and development.

The information contained in BRW is updated concurrently with that in the individual databases.

Resource research circulation

If a researcher obtains good results with NBRP resources, such a result may be useful for later researchers who use the same resources. We therefore ask researchers to feed back information on papers that report the results of their studies, and we also collect other papers in which NBRP resources are used. This information is used to create an open database (RRC) of papers related to the NBRP resources. RRC also provides an online registration system through which a paper can be submitted merely by entering its PubMed identification number and the name of the resource used in the paper. We have asked many researchers to use this system to feed back on the papers that they have published. It would be ideal for experimental researchers if, as in the case of the accession numbers of the DNA Data Bank, we could establish a system in which detailed information on the bioresources described in 'Materials and Methods' sections of papers could be obtained from public databases and, furthermore, the materials themselves could be easily acquired.

The NBRP databases

Table 1 shows the names and features of the NBRP component databases, organized by organism group. As shown in Table 1, the NBRP resource collection covers

a wide range of taxa with indications of the presence or absence of the following collections: (i) naturally mated lines (including wild species, cultivated species, inbred lines and spontaneous mutants); (ii) genetically engineered organisms (including induced mutants, transgenic strains, transposon-inserted strains, deletion strains, consomic lines, genome-wide knockout strains, and enhancer trap lines); and (iii) DNA (plasmids, vectors and genomic/cDNA clones). The numbers of records of resources vary according to the group of organisms, and the resources in classification (3) account for 95% of the total. The External DB column in Table 1 indicates databases that have one-way links or cross links to external comprehensive databases for the model organisms. The Collaborating DB inside the parenthesis indicates an external database with which the NBRP database collaborates.

Types and names (identifiers) of the resources, and distribution/deposition methods (such as MTA) are essential information for all resources. For DNA clones, the accession numbers in the DNA Data Bank, together with homologous sequences, are the items of information common to all organism groups. NBRP also includes some activities in which the distribution of resources requires a prescribed review procedure (such as the Human ES-cells and Macaque). Activities where resources cannot be obtained by proliferation, but require non-invasive sampling or sampling from dead animals, are provided only for research that meets prescribed standards [such as the Great Ape Information Network (GAIN)]. The Global Biodiversity Information Facility (GBIF; http://www.gbif.org/) is an activity that forms the Japanese node of an international project that is involved in collecting and creating a database on specimens and observation data held in museums. Information from the GBIF and resources from the GAIN and Macaque databases are generally not distributed, so that they are not included in BRW.

As illustrative examples of the features described above, the databases for rice, *Escherichia coli*, *Caenorhabditi elegans* and the rat are described in more detail below, and some details of the RIKEN BioResource Center databases are also given.

NBRP-rice (Oryzabase) (25). Oryzabase, a comprehensive database that is currently used by many researchers, succeeded the Database of Resources and Trait Genes, which was launched in 1995. Genomic information has been added to the database on resources and genes so that it has become comprehensive and now functions as the NBRP database. The characteristic wild-rice collection consists of three different core collections. Core collection Rank 1 relates to 18 species from nine genomes and contains detailed phenotype data and many photographs. The collection of unique mutant strains includes strains for which trait genes have been identified, as well as strains that have phenotypes for which the responsible genes are not known. The former set of strains can be retrieved on the basis of genetic information, and the latter set of strains can be retrieved on the basis of the phenotype. The data on both sets of strains include photographs.

We constantly update the dictionary of trait genes with data extracted from published papers, and we have also been promoting the establishment of correspondences between trait genes and the accession numbers of DNA sequences, ORF numbers of genome projects and displays of chromosomes by using physical maps. Oryzabase also provides access to a Web site for online submission of new rice genes, which is based on the Gene Nomenclature System for Rice as determined by the Committee on Gene Symbolization, Nomenclature and Linkage of the Rice Genetics Cooperative (CGSNL) (26). Users can give feedback on an individual gene through the detailed pages for that gene.

Basic information on rice, including the definitions of the tissue-specific developmental stages (such as embryo, inflorescence, leaf, root, anther, ovule, pollen mother cell, stoma or vascular bundle) and information on tissue-specific or developmental stage-specific gene expression are also available through this web site.

A version of Textpresso, a text-mining system for scientific papers developed by the Generic Model Organism Database and adapted for use with rice, has been constructed and is available through Oryzabase. It provides access to a total of ~20 000 rice-related papers (abstracts and titles) in PubMed, and is linked from the home page of Textpresso.

Oryzabase has links to two external databases: Gramene and IRRI. Genetic information is linked to the former database, and wild-strain information is linked to the latter. Oryzabase also allows BLAST searches to be made through all genomes or by chromosomes, provides a tool for extracting the specified region of genome sequences, and can provide downloadable text files of almost all information.

NBRP-E. coli. Of all the organisms in the NBRP collection, *E. coli* was the first to have its genome sequence determined. A feature of NBRP-*E. coli* is its genome-wide genetically engineered strain collections. At the web site of NBRP-*E. coli* (http://www.shigen.nig.ac.jp/ecoli/strain/), users can browse outlines and lists of collections and search full data or perform queries by specifying details. Although online requests are available, these often take time because of the complexity of the MTA process. For the convenience of users, a tracking system has been introduced that allows users to check the progress of their requests in real time.

The information center maintains profiling of *E. coli* chromosomes (PEC), an information site of a project on essential *E. coli* genes (27). Through PEC, the information center has made genome maps and genetic information available. The resource information contained in NBRP-*E. coli* is cross linked with PEC to permit access to resources through the maps. For example, by clicking the deletion regions of extensive deletion mutants, or by clicking the gene parts of mutants mutated by gene units, the user can access detailed pages that are linked to request pages for the resources. PEC has many links to external databases, including NCBI, UniProtKB, COG,

Table 1. The NBRP databases

	Database name	Taxon	Wild/ inbred/ landrace	Mutants	Plasmid/ vector/ clones	Map (Physical or Linkage)	Gene (mutated genes/all genes)	Phenotype	SNPs	Blast service	Images	External DB (collaborating DB)[reference]	Featured contents/ services
1	NBRP-Macaque	Vertebrata-mammalia-primate	o		o								
2	NBRP-Mouse (Riken BRC)	Vertebrata-mammalia-murinae-mus	o	o	o	P	M	o	o	o	o	MGI[3], (IMSR, JMSR;http://www.shigen.nig.ac.jp/mouse/jmsr/)	
3	NBRP-Rat	Mammalia-murinae-rattus	o	o	o	P	M	o	o	o	o	RGD[4], (JMSR)	graphical display of the phenotype data
4	NBRP-Xenopus	Vertebrata-amphibia	o		o					o			
5	NBRP-Zebrafish	Vertebrata-actinopterygii-cypriniformes-danio		o			M	o			o	ZFIN[5]	
6	NBRP-Medaka	Vertebrata-actinopterygii-beloniformes-oryzias	o	o	o	P	M	o		o	o		atlas, phylogenetic tree
7	NBRP-Ciona	Ascidiacea-ciona		o	o		M	o			o	(Ghost,[6] CIPRO;http://cipro.ibio.jp/new/)	
8	NBRP-Drosophila (DGRC)	Arthropoda-insecta-diptera-drosophilidae	o	o	o	P	A	o	o	o	o	FlyBase[7]	
9	NBRP-Silkworm (SilkwormBase)	Arthropoda-insecta-lepidoptera-bombyx		o		L	M	o			o		laeva period time
10	NBRP-*C. elegans*	Pseudocoelomata-nematoda		o			M					Wormbase[8]	
11	NBRP-Rice (Oryzabase)	Viridiplantae-poaceae-oryzeae	o	o		P,L	A	o		o	o	Gramene[9], IRRI; http://beta.irri.org/index.php/Home/Welcome/Frontpage.html	developmental stage
12	NBRP-Barley	Viridiplantae-poaceae-triticeae-hordeum	o		o	L		o		o	o		phenotype data
13	NBRP-Wheat (KOMUGI)	Viridiplantae-poaceae-triticeae	o	o	o	L	A	o		o	o	(TriFLDB[10])	gene catalogue
14	NBRP-*Arabidopsis* (Riken BRC)	Viridiplantae-brassicales-arabidopsis		o	o			o		o	o		
15	NBRP-*Chrysanthemum*	Viridiplantae-asterids-chrysanthemum	o					o			o		
16	NBRP-Morning glory	Viridiplantae-solanales-ipomoeeae		o	o	L	M	o			o		
17	NBRP-Lotus/ Glycine (LegumeBase)	Fabales-fabaceae	o	o	o	L	M	o			o	(miyakogusa.jp[11], Soybean Full-length cDNA database[12])	
18	NBRP-Tomato	Viridiplantae-solanales-lycopersicon			o							(KafTom;http://www.pgb.kazusa.or.jp/kaftom/, MiBASE[13])	
19	NBRP-Algae	14 phyla (eukaryota and bacteria)	o					o			o		photograph, phylogenetic tree
20	NBRP-Yeast	Fungi-ascomycota		o	o	P	A					SGD[14], geneDB[15]	
21	NBRP-Cellular slime mold	Mycetozoa-dictyostelliida			o							dictyBase[16]	
22	NBRP-Prokaryote *E. coli*	Bacteria-proteobacteria		o	o	P	A			o		PEC[17], EcoGene[18], EcoCyc[19], COG[20], NCBI[1], SwissProt[21], GTOP[22], KEGG[23], InterPro[24]	
23	NBRP-Prokaryote *B. subtilis*	Bacteria-firmicutes-bacilli-bacillaceae		o			M						
24	NBRP-Pathogenic microbes	Bacteria and protozoa	o					o			o		
25	NBRP-General microbes (Riken BRC, JCM)	Bacteria	o										

EcoCyc, GTOP, EcoGene and KEGG. We also perform similarity searches on gene sequences to add domain information from Pfam and PROSITE. In addition, PEC provides BLAST search through two different strains, MG1655 and W3110; all ORF, and essential genes. The tool for specifying the desired fragments of genome sequences is also available. Almost all the information in the database can be downloaded in the form of text files.

NBRP-C. elegans. NBRP-*C. elegans* (http://www.shigen.nig.ac.jp/c.elegans/) is a smaller database than the two databases discussed above. Each record consists only of the allele of the deletion strain, a systematic identification tag for the gene (the CGC name), information on the position of the gene on the chromosome, information on the positions of deletion regions, and primer information. However, the CGC name, allele and sequence are cross linked to the corresponding page of WormBase, a comprehensive database on *C. elegans*, providing an easy access to that database. In the case of this resource, requested mutants are isolated from the pool after the request. Users can check on the state of progress of isolation online. Mutants that have been isolated once will be listed as available mutants (isolated), so that other users can request the same mutants. NBRP-*C. elegans* is unique in that information on researchers who have received mutants from this project is made public online. Because almost all papers in which the resources are used contain the names of these resources, and the information is fed back from paper-registration sites, the *C. elegans* database automatically reflects information in the RRC. This is a model case of good circulation between resources and researches.

NBRP-rat. Because the database of rat resources (http://www.anim.med.kyoto-u.ac.jp/NBR/) contains substantial characterization data from individual resources and has many tools for browsing this data, it is efficient in allowing researchers to find the resources best suited to their research from a range of trait information. For example, the top page displays a pie chart that shows a breakdown of the research fields in which the resources are used, and by clicking a research field of interest, the user can obtain a list of resources related to that particular research field. Another example, 'Phenome Project' provides 109 items of physiological, behavioral and anatomical phenome data in nine tables and in strain-distribution maps with the two items selected by the user as the abscissa and ordinate axes. Users can access detailed pages for strains by clicking data points in the maps. 'Genome' provides access to polymorphism data for 357 simple sequence-length polymorphism (SSLP) markers, obtained from investigations on more than 150 strains. A phylogenetic tree, constructed from polymorphism data, is also available, through which users can access detailed pages on resources.

Besides characterization data, the database also contains detailed pages on resources, including the preservation status, genetic status, research category, origin, genotype, references and a link to the Rat Genome Database (RGD). Recent enhancements include a BAC browser for F344 and LE BAC end sequences, and the addition of functional polymorphism data obtained by comparing 16 disease-associated gene mutations among multiple strains. The ENU-induced mutant archive (28) is also provided at this site.

NBRP-RIKEN BRC

The RIKEN BioResource Center (BRC) was established in 2001 with the aim of becoming the finest core bioresource facility in the world. Since then, it has been engaged in collecting bioresources developed mainly by Japanese scientists. These bioresources include living strains of mice and *Arabidopsis*, human and animal cells, DNA materials and various microbes for which the RIKEN BRC has been designated the national core facility by the NBRP. The RIKEN BRC preserves these bioresources under conditions of strict quality control for provision to the scientific community. The RIKEN BRC also collects information on the whereabouts and characteristics of the bioresources, constructs databases, and offers the bioresource information to the research community within and outside Japan.

The Animal Search System (http://www2.brc.riken.jp/lab/animal/search.php) allows keyword searches to be performed on >2000 strains of mice available from the RIKEN BRC, including transgenic, knockout, inbred, wild-derived, ENU mutant and congenic strains. The system also provides detailed information on each strain, such as the strain name, description, gene details, references, availability status, health report, depositor, specific terms and conditions for distribution, and image(s), which capture the characteristics of the strain. Some gene symbols have links to the Mouse Genome Informatics (MGI) database, which leads to further detailed information on the particular gene. As well as being publicly available, the up-to-date information on mice is also sent to the International Mouse Strain Resource (IMSR, http://www.findmice.org/) database, to which the RIKEN BRC is a contributing repository. The RIKEN BRC Mouse Phenome Database (RMPD) (http://www.brc.riken.jp/rmpd/mouse_phenome_top.html) contains phenotypic data on the physiology, biochemistry, hematology and morphology of inbred, mutant, wild-derived and recombinant inbred strains of mice, and enables biomedical researchers to find appropriate strains for their researches.

The *Arabidopsis* transposon tagged lines can be searched in a web-based catalogue (http://www.brc.riken.jp/lab/epd/catalog/transposon.html), where information on insert positions of transposons and adjacent genes can be obtained for more than 15 000 lines. The SENDAI *Arabidopsis* Seed Stock Center (SASSC) database (http://www.brc.riken.jp/lab/epd/SASSC/) provides information on its collection (wild type and mutant), such as strain name, region of collection and phenotypic remarks. The RIKEN *Arabidopsis* full-length cDNA (RAFL) clone database (http://www.brc.riken.jp/lab/epd/catalog/cdnaclone.html) contains more than 250 000 clones and the users can retrieve clones by

NCBI accession number, AGI code, clone name or sequence homology. The Systematic Consolidation of *Arabidopsis* and other Botanical REsources (SABRE) database (http://saber.epd.brc.riken.jp/sabre/SABRE0101.cgi) offers searches of BRC plant DNA resources across the species that it contains.

The database of human and animal cells (http://www2.brc.riken.jp/lab/cell/search.php) has its origins in the database of the RIKEN Cell Bank, which began collecting and distributing cell resources in 1987. Starting from a stand-alone database for internal use, it has been transformed to allow searching of cell resources and the provision of the associated information through the Internet. Because the Cell Bank began its activity as a division of the BRC, its stock list has grown with the addition of various new kinds of cell resources, such as Epstein–Barr virus-transformed B cell lines, human somatic stem cells, embryonic stem cell lines and induced pluripotent stem cell lines. The function of the database has accordingly been enhanced to allow the presentation of other items of information that differ from resource to resource. The information available therefore depends on the resource; for example, information on the origin, morphology, culture conditions, restrictions on distribution, results of short tandem repeat analyses, images, etc. can be obtained for a conventional cell line resources.

Before the BRC was established, DNA materials were collected and distributed as an activity of the RIKEN DNA Bank. A keyword search system (http://www.brc.riken.jp/lab/dna/search/index.html) is available for DNA clones, vectors and recombinant adenoviruses. Resources are retrieved by plasmid name, by gene name or symbol, or by accession number. In the Geneset Bank database (http://www.brc.riken.jp/lab/dna/en/GENESETBANK/index.html), more than 20 illustrations for principal gene pathways are implemented, and DNA materials can be easily found according to the gene pathways to which they belong.

The microbial resource collection in RIKEN was founded in 1980 as the Japan Collection of Microorganisms (JCM), and the construction of a database for the collection began at its inception. The Web-based online catalog database (http://www.jcm.riken.jp/JCM/catalogue.shtml) was launched in 1995, and it has since been improved and updated. It now provides access to information on more than 11 000 available strains, which are searchable by their accession number, scientific name and keywords on strain data. It is also possible to search for JCM strains that are equivalent to those in other culture collections. The database contains various useful items of information about strains, such as culture media and conditions, history (including isolation source), taxonomic data and references. The DNA sequence data, critical for the phylogeny and the genome data are linked to the DDBJ database for each strain.

Future directions

We will continue to improve the content of individual databases and upgrade the functions available at the integrated database-retrieval site. We also hope to expand external access to the databases and expand collaboration with other databases to permit access to our resources by a wider range of users. In particular, we will examine a new possibility for interconnecting reference data and resource databases with the aim of construction a virtual international network.

FUNDING

Ministry of Education, Culture, Sports, Science and Technology (MEXT) (to National Bio Resource Project). Funding for open access charge: The Management Expenses Grant for National University Cooperation, MEXT.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
2. Sugawara,H., Ikeo,K., Fukuchi,S., Gojobori,T. and Tateno,Y. (2009) DDBJ dealing with mass data produced by the second generation sequencer. *Nucleic Acids Res.*, **37**, D16–D18.
3. Blake,J.A., Bult,C.J., Eppig,J.T., Kadin,J.A., Richardson,J.E. and the Mouse Genome Database Group. (2009) The Mouse Genome Database genotypes:phenotypes. *Nucleic Acids Res.*, **37**, D712–D719.
4. Dwinell,M.R., Worthey,E.A., Shimoyama,M., Bakir-Gungor,B., DePons,J., Laulederkind,S., Lowry,T., Nigram,R., Petri,V., Smith,J. *et al.* (2009) The Rat Genome Database 2009: variation, ontologies and pathways. *Nucleic Acids Res.*, **37**, D744–D749.
5. Sprague,J., Bayraktaroglu,L., Bradford,Y., Conlin,T., Dunn,N., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Knight,J. *et al.* (2008) The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. *Nucleic Acids Res.*, **36**, D768–D772.
6. Satou,Y., Kawashima,T., Shoguchi,E., Nakayama,A. and Satoh,N. (2005) An integrated database of the ascidian, Ciona intestinalis: towards functional genomics. *Zoolog Sci.*, **22**, 837–843.
7. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al.* (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.*, **37**, D555–D559.
8. Rogers,A., Antoshechkin,I., Bieri,T., Blasiar,D., Bastiani,C., Canaran,P., Chan,J., Chen,W.J., Davis,P., Fernandes,J. *et al.* (2008) WormBase2007. *Nucleic Acids Res.*, **36**, D612–D617.
9. Liang,C., Jaiswal,P., Hebbard,C., Avraham,S., Buckler,E.S., Casstevens,T., Hurwitz,B., McCouch,S., Ni,J., Pujar,A. *et al.* (2008) Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res.*, **36**, D947–D953.
10. Mochida,K., Yoshida,T., Sakurai,T., Ogihara,Y. and Shinozaki,K. (2009) TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. *Plant Physiol.*, **150**, 1135–1146.
11. Sato,S., Nakamura,Y., Kaneko,T., Asamizu,E., Kato,T., Nakao,M., Sasamoto,S., Watanabe,A., Ono,A., Kawashima,K. *et al.* (2008) Genome structure of the legume, *Lotus japonicus*. *DNA Res.*, **15**, 227–239.
12. Umezawa,T., Sakurai,T., Totoki,Y., Toyoda,A., Seki,M., Ishiwata,A., Akiyama,K., Kurotani,A., Yoshida,T., Mochida,K. *et al.* (2008) Sequencing and analysis of approximately 40,000 soybean cDNA clones from a full-length-enriched cDNA library. *DNA Res.*, **15**, 333–346.
13. Yano,K., Watanabe,M., Yamamoto,N., Tsugane,T., Aoki,K., Sakurai,N. and Shibata,D. (2006) MiBASE: a database of a miniature tomato cultivar Micro-Tom. *Plant Biotechnol.*, **23**, 195–198.
14. Hong,E.L., Balakrishnan,R., Dong,Q., Christie,K.R., Park,J., Binkley,G., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G. *et al.* (2008) Gene Ontology annotations at SGD: new data

sources and annotation methods. *Nucleic Acids Res.*, **36**, D577–D581.
15. Hertz-Fowler,C., Peacock,C.S., Wood,V., Aslett,M., Kerhornou,A., Mooney,P., Tivey,A., Berriman,M., Hall,N., Rutherford,K. *et al.* (2004) GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res.*, **32**, D339–D343.
16. Fey,P., Gaudet,P., Curk,T., Zupan,B., Just,E.M., Basu,S., Merchant,S.N., Bushmanova,Y.A., Shaulsky,G., Kibbe,W.A. *et al.* (2009) dictyBase—a Dictyostelium bioinformatics resource update. *Nucleic Acids Res.*, **37**, D515–D519.
17. Yamazaki,Y., Niki,H. and Kato,J. (2008) Profiling of Escherichia coli Chromosome Database. *Methods Mol. Biol.*, **416**, 385–389.
18. Rudd,K.E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 60–64.
19. Keseler,I.M., Bonavides-Martínez,C., Collado-Vides,J., Gama-Castro,S., Gunsalus,R.P., Johnson,D.A., Krummenacker,M., Nolan,L.M., Paley,S., Paulsen,I.T. *et al.* (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res.*, **37**, D464–D470.
20. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedovm,S.I., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
21. UniProt Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
22. Fukuchi,S., Homma,K., Sakamoto,S., Sugawara,H., Tateno,Y., Gojobori,T. and Nishikawa,K. (2009) The GTOP database in 2009: updated content and novel features to expand and deepen insights into protein structures and functions. *Nucleic Acids Res.*, **37**, D333–D337.
23. Kanehisa,M., Ataki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
24. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database (2009). *Nucleic Acids Res.*, **37**, D224–D228.
25. Yamazaki,Y. and Kurata,N. (2006) Oryzabase, An Integrated Biological and Genome Information Database for Rice. *Plant Physiol.*, **140**, 12–17.
26. Susan,R.M. and CGSNL (Committee on Gene Symbolization, Nomenclature and Linkage, Rice Genetics Cooperative. (2008) Gene nomenclature system for rice. *Rice*, **1**, 72–84.
27. Kato,J. and Hashimoto,M. (2008) Construction of long chromosomal deletion mutants of *Escherichia coli* and minimization of the genome. *Methods Mol. Biol.*, 279–293.
28. Mashimo,T., Yanagihara,K., Tokuda,S., Voigt,B., Takizawa,A., Nakajima,R., Kato,M., Hirabayashi,M., Kuramoto,T. and Serikawa,T. (2008) An ENU-iuduced mutant archive for gene targeting in rats. *Nat. Genet.*, **40**, 514–515.

Biochemical and Biophysical Research Communications 397 (2010) 697–705

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Gene expression profiles of cryopreserved CD34$^+$ human umbilical cord blood cells are related to their bone marrow reconstitution abilities in mouse xenografts

Kazuhiro Sudo [a,1], Jun Yasuda [b,c,*,1], Yukio Nakamura [a,**]

[a] Cell Engineering Division, RIKEN BioResource Center, Tsukuba, Japan
[b] Omics Science Center, RIKEN, Yokohama, Japan
[c] Department of Cell Biology, The JFCR-Cancer Institute, Japan

ARTICLE INFO

Article history:
Received 1 June 2010
Available online 4 June 2010

Keywords:
Gene expression profile
Transplantation
Umbilical cord blood

ABSTRACT

Human umbilical cord blood (UCB) cells are an alternative source of hematopoietic stem cells for treatment of leukemia and other diseases. It is very difficult to assess the quality of UCB cells in the clinical situation. Here, we sought to assess the quality of UCB cells by transplantation to immunodeficient mice. Cryopreserved CD34$^+$ UCB cells from twelve different human donors were transplanted into sublethally irradiated NOD/shi-scid Jic mice. In parallel, the gene expression profiles of the UCB cells were determined from oligonucleotide microarrays. UCB cells from three donors failed to establish an engraftment in the host mice, while the other nine succeeded to various extents. Gene expression profiling indicated that 71 genes, including *HOXB4*, *C/EBP-β*, and *ETS2*, were specifically overexpressed and 23 genes were suppressed more than 2-fold in the successful UCB cells compared to those that failed. Functional annotation revealed that cell growth and cell cycle regulators were more abundant in the successful UCB cells. Our results suggest that hematopoietic ability may vary among cryopreserved UCB cells and that this ability can be distinguished by profiling expression of certain sets of genes.

1. Introduction

Human UCBs are an important source of hematopoietic stem cells for treatment of various hematological disorders [1] and malignant diseases [2]. UCBs have the advantages of a more relaxed histocompatibility requirement than bone marrow cells and of the absence of any risk to the donor in their collection. These advantages make UCBs a versatile option for bone marrow reconstitution therapy [3]. The clinical outcome of UCB transplantation is comparable to allogeneic bone-marrow transplantation [4,5].

However, it has been found that 10–20% of patients experience insufficient engraftment of UCB donor cells and that some cases may have no evidence of bone marrow rescue [4–8]. Success of UCB transplantation is associated with various factors, such as use of low-dose total-body irradiation, patient age, and less severe disease progression [5,9]. In addition to patient-related factors, it is also possible that the quality of the stem cells in the UCB, such as cell viability, pluripotency, and resistance to stress, affects the outcome of the allograft. Currently, the only available information on the influence of the quality of UCBs and outcome of the allograft is that a high dose of CD34$^+$ cells is required to ensure successful engraftment of transplanted UCBs [5,8,10]. One of the particular difficulties for assessing the influence of the quality of the UCBs on the outcome of therapeutic use is that the clinical condition of the recipients is so variable as to preclude any direct comparison of the quality of UCB allografts. Several factors, such as ethnicity, birth weight, sex, and type of delivery, are correlated with the ability of cryopreserved human UCBs to form colonies in *in vitro* culture [11].

Stevens et al. reported that the numbers of nucleated red blood cells are correlated with those of hematopoietic progenitor cells in human UCBs; moreover, UCBs with a high count of nucleated red blood cells show faster engraftment [12]. To date, however, no molecular indicators have been identified for assessing the quality of human UCBs. In this study, we used a bone marrow reconstitution assay to assess xenograft success in mice, and thus enable evaluation of the quality of human cryopreserved UCBs. Additionally, we performed gene expression profiling of these UCBs in order to identify possible molecular markers for quality assessment. For the former assay, we used sublethally irradiated immunodeficient NOD/shi-scid Jic (NOD/SCID) mice as they offer reproducible recipient conditions and are therefore ideal for investigating potential

Abbreviations: UCB, umbilical cord blood; NOD/SCID, non-obese diabetes/severe combined immune-deficient.

* Corresponding author at: Department of Cell Biology, The JFCR-Cancer Institute, Ariake 3-8-31, Koto-ku, Tokyo 135-8550, Japan. Fax: +81 3 3570 0475.
** Corresponding author. Address: Cell Engineering Division, RIKEN BioResource Center, Koyadai 3-1-1, Tsukuba, Ibaraki 305-0074, Japan. Fax: +81 29 836 9049.
E-mail addresses: yasuda-jun@umin.ac.jp (J. Yasuda), yukionak@brc.riken.jp (Y. Nakamura).

[1] These two authors contributed equally to this work.

0006-291X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2010.06.010

differences in hematopoietic reconstitution abilities among human UCBs [13,14].

The NOD/SCID mouse system was first exploited by Gan et al. to compare human hematopoietic stem cells transplanted immediately after collection with those subjected to *ex vivo* culture with stromal cells; they found a decrease in the rate of repopulating cells after transplantation of cultured hematopoietic cells [13]. We transplanted UCBs into NOD/SCID mice and analyzed the hematopoietic cells present in peripheral blood and the bone marrow. We also determined the gene expression profiles of the $CD34^+$ UCB cells in their pre-transplantation condition, and searched for an expression signature that correlated with the success of transplantation of the $CD34^+$ UCB cells.

2. Materials and methods

2.1. CD34-positive cells

$CD34^+$ UCB cells were obtained from the Stem Cell Resource Network in Japan (Banks at Miyagi, Tokyo, Kanagawa, Aichi, and Hyogo) through the RIKEN BioResource Center (Tsukuba, Ibaraki, Japan).

2.2. Mice

Seven-week-old female NOD/SCID mice were purchased from CLEA Japan (Tokyo, Japan). The mice were used within two weeks of delivery. Four to six hours prior to cell transplantation, the mice were given a 300 cGy dose of γ-rays.

2.3. Transplantation assay

$CD34^+$ cells (3×10^5 cells) from each sample of UCB were suspended in 600 μl MEM-α containing 10% fetal bovine serum (FBS); 200 μl of the suspension (1×10^5 cells) was then injected into the tail vein of each of three NOD/SCID mice. This procedure was repeated for each of the 12 UCB samples.

2.4. Flow cytometry

Twelve weeks after transplantation, peripheral blood samples were obtained from the retro-orbital venous plexus, and bone marrow cells were obtained after sacrifice. The peripheral blood and bone marrow cells were stained with monoclonal antibodies (MoAbs) and analyzed by FACS Calibur (BD Biosciences, San Jose, CA, USA). The red blood cells in the peripheral blood samples were lysed using red blood cell lysis buffer (140 mM NaCl, 1 mM $NaHCO_3$) prior to cell staining. The following MoAbs were purchased from BD Biosciences: a fluorescein isothiocyanate (FITC)-conjugated MoAb against human CD45 (CD45-FITC), a phycoerythrin (PE)-conjugated MoAb against human CD34 (CD34-PE), an allophycocyanin (APC)-conjugated MoAb against mouse CD45 (mCD45-APC), CD19-PE, CD33-APC, Glycophorin A-FITC, and TER119-PE. Cell viability was determined after propidium iodide (SIGMA, St Louis, MO, USA) staining. Data from $1x10^4$ living cells were collected and analyzed using CellQuest Pro (BD Biosciences) and FlowJo (Tree Star Inc., Ashland, OR, USA) analysis software. The rate of chimerism (%) was calculated from the flow cytometry data as follows: rate of chimerism of human cells (%) = [% human $CD45^+$ cells/(% human $CD45^+$ cells + % mouse $CD45^+$ cells)] × 100.

2.5. Oligonucleotide microarray analysis

Total RNA was extracted using the RNeasy mini kit (Qiagen, Hilden, Germany) from an aliquot of each sample of human cryopreserved $CD34^+$ UCB cells at the time of thawing. Then, 250 ng of each total RNA was subjected to reverse transcription and isothermal linear amplification using Ribo-SPIA (NuGEN, San Carlos, CA, USA) [15], using a modification of the manufacturer's recommended protocol. The linearly amplified cDNAs served as templates for the *in vitro* transcription generating hybridization target cRNAs using the Low RNA Fluorescent Linear Amplification Kit Plus (Agilent, Santa Clara, CA, USA). The amplified cRNAs were labeled with Cy3 dye and used for hybridization to the oligonucleotide microarray (Agilent human whole genome 4 × 44) following the manufacturer's protocols. Hybridization signals were scanned with the Agilent Technologies Scanner G2505C (Agilent) and were extracted from the scanned images by the use of Feature Extraction Ver. 9.5.3 (Agilent). All microarray data reported in this paper is described in accordance with MIAME guidelines and the data has been deposited in the GEO (Gene Expression Omnibus) database at the National Center for Biological Information, National Institute of Health (USA). The accession number for the dataset is GSE19835.

2.6. Normalization of gene expression profiles

Quality control and array normalization was performed in the R statistical environment (http://www.r-project.org) using the Agi4x44PreProcess package downloaded from the Bioconductor web site (http://bioconductor.org/). The data files were appropriately edited with text editing software to render the files compatible for the Agi4x44PreProcess packages. The normalization and filtering steps were based on those described in the Agi4x44PreProcess reference manual.

2.7. Statistical analyses

A heat map of differentially expressed genes was generated using Gene Cluster 3.0 software [16] and visualized with TreeView software [17]. Overexpressed genes specific for each phenotype were identified by Student's *t*-tests, and potential false-positives were removed by the Benjamini–Hochberg method. The GSEA analysis [18] was carried out using GSEA Java desktop software (version 2.04, http://www.broadinstitute.org/gsea/). The C2 curated gene set and C2-all gene set for GSEA analysis were retrieved from the Molecular Signatures Database (MSigDB: http://www.broadinstitute.org/gsea/downloads.jsp). The DAVID functional annotation system was used as described by Huang et al. [19].

2.8. Quantification of RNAs using real-time PCR

Real-time PCR was performed using the ABI 9500 Real-time PCR system (Applied Biosystems, Foster City, CA, USA) and each amplification reaction was performed in quadruplicate. For quantification of miRNAs, the cDNAs prepared for the microarray analysis were used as templates and PCR was performed with SYBR premix Ex Taq (Perfect Real Time: Takara Bio Inc., Shiga, Japan) with GAPDH as the loading control. The nucleotide sequences used for PCR amplification are given in *Supplementary* Table 1.

3. Results

3.1. Hematopoiesis reconstitution in immunodeficient mice xenografted with $CD34^+$ cells

We first compared the relative abilities of 12 samples of $CD34^+$ UCB cells from different donors to form engraftments in immunodeficient NOD/SCID mice. Each animal was injected with 1×10^5 cells; the cells from each donor were injected into three mice and the remaining cells were used for the microarray gene expression

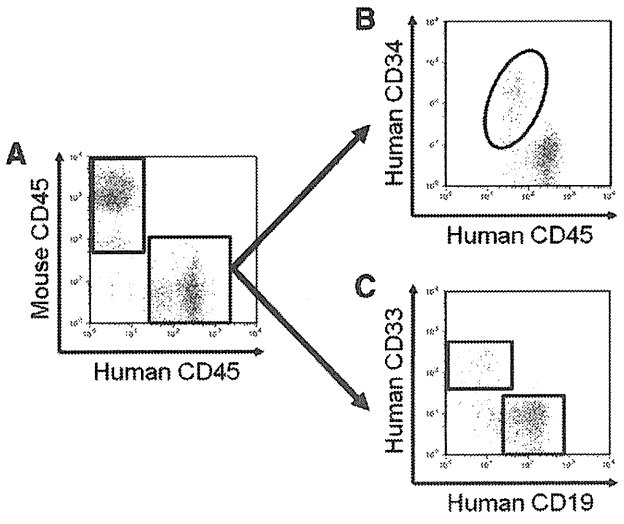

Fig. 1. Classification of human and mouse hematopoietic cells. (A) Separation of human and mouse hematopoietic cells from the bone marrow of a mouse xenograft by flow cytometry. The horizontal axis and the vertical axis indicate the signal intensities for human CD45- and mouse CD45-specific antibodies, respectively. (B) Quantification of human $CD34^+$ cells among human $CD45^+$ cells. (C) Quantification of human $CD19^+$ and $CD33^+$ hematopoietic cells among human $CD45^+$ cells. (A–C) Examples of gating are indicated.

analysis. Twelve weeks after cell transplantation, peripheral blood cells were collected from each mouse. After collection of peripheral blood, the mice were sacrificed and bone marrow cells were obtained. The presence and relative numbers of human hematopoietic cells in the peripheral blood and bone marrow were determined by flow cytometry (Fig. 1). The rate of chimerism for human hematopoietic cells was calculated as the proportion of human $CD45^+$ cells in all leukocytes in the peripheral blood and bone marrow, respectively (Fig. 1A and Section 2).

Although all of the 12 $CD34^+$ UCB samples produced human $CD45^+$ cells in at least one of the three recipient mice (>0.1% chimerism), three samples (H07041, H07056, and H07112) produced very small numbers of human $CD45^+$ cells in both the peripheral blood and bone marrow of all recipient mice; we designated these three samples as "failed UCBs" (Fig. 2). The other nine $CD34^+$ UCB samples established obvious engraftments in at least one of the recipient mice; we designated these nine samples as "successful UCBs" (Fig. 2). However, the frequency of human $CD45^+$ cells varied among the successful UCBs. For example, UCB H07088 and UCB H07133 produced rates of 6.88–35.8% and 90.0–94.3%, respectively, in the bone marrow of the mice (Fig. 2, *Supplementary* Table 2). The frequencies of $CD34^+$, $CD19^+$ and $CD33^+$ cells in the human $CD45^+$ cells were also calculated for peripheral blood and bone marrow samples that had more than 3% human $CD45^+$ cell chimerism (Fig. 1B and C, and *Supplementary* Table 2). The proportions of lymphoid and myeloid cells among the human $CD45^+$ cells in the peripheral blood and bone marrow of mice that received one of the nine successful UCBs also varied (*Supplementary* Table 2).

3.2. Gene expression profiles of the 12 human UCB samples

We sought to determine if there was any connection between the gene expression profiles of the UCBs and their abilities to achieve successful engraftment. Gene expression profiles were determined by oligonucleotide microarray analyses using total RNAs from the UCB cells. The amount of total RNA obtained from each aliquot (generally several nanograms) was insufficient to perform the assay without a further *in vitro* amplification step. After amplification of the cDNAs, complementary RNAs were used for hybridization. Probes that were positive in 75% of all samples were selected for further analyses; in total, 23,807 of the 45,015 probes tested were selected.

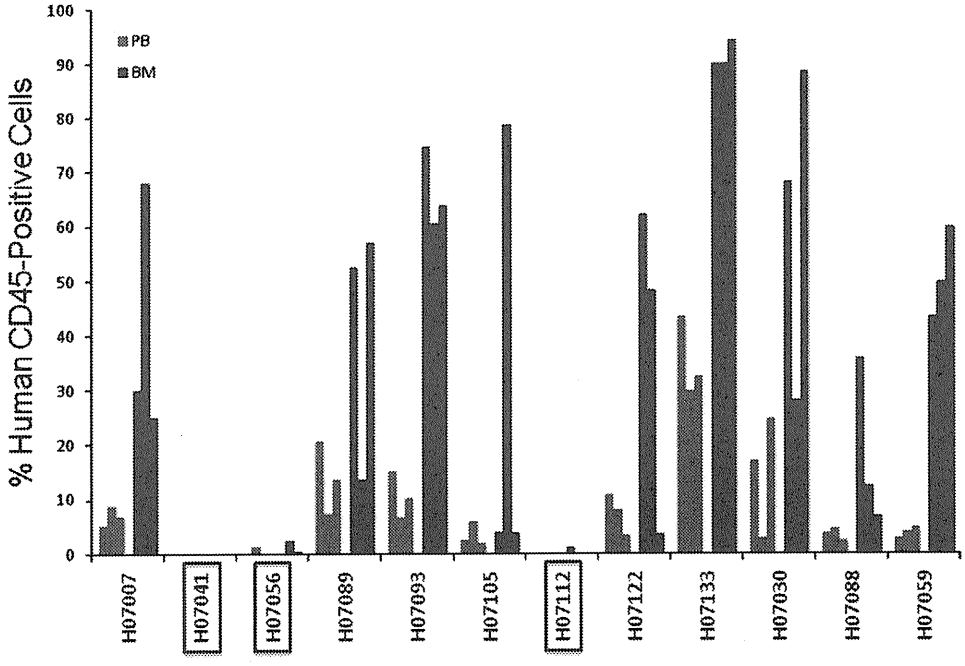

Fig. 2. Chimerism in irradiated NOD/SCID mice following injection of cryopreserved human $CD34^+$ UCB cells. The bar graphs indicate the percentages of human $CD45^+$ cells (see Fig. 1A and Section 2) from either the bone marrow (purple bar) or the peripheral blood (blue bar) of mouse xenografts. Each bar represents a single mouse. The names under the horizontal axis indicate the different human UCB samples and the boxed names indicate those that failed to engraft (see text).

Table 1
Genes that are overexpressed in successful or failed UCBs.

Gene Symbol	Accession Number	ENTREZ ID	Category	*t*-Test	Fold change
Overexpressed in successful UCBs					
DNHD1	AK074178	144132	Dynein heavy chain domain 1	0.003748741	4.754
HOXB4	NM_024015	3214	Homeobox B4	0.001103195	4.548
SMC1A	NM_006306	8243	Structural maintenance of chromosomes 1A	0.000781411	4.378
MED1	BC060758	5469	Mediator complex subunit 1	0.00311929	4.041
SNRNP48	NM_152551	154007	Small nuclear ribonucleoprotein 48 kDa (U11/U12)	4.04958E-05	3.252
ZNF12	NM_016265	7559	Zinc finger protein 12	0.000123475	3.079
CEBPB	NM_005194	1051	CCAAT/enhancer binding protein (C/EBP), beta	0.000592075	3.064
CASKIN1	NM_020764	57524	CASK interacting protein 1	7.18248E-05	3.039
CYFIP2	NM_001037332	26999	Cytoplasmic FMR1 interacting protein 2	0.000128743	2.849
IL17D	NM_138284	53342	Interleukin 17D	8.5289E-05	2.827
HIST2H2AB	NM_175065	317772	Histone cluster 2, H2ab	0.0018657	2.820
C9orf102	NM_020207	375748	Chromosome 9 open reading frame 102	0.004144549	2.788
ZNF331	NM_018555	55422	Zinc finger protein 331	0.000439891	2.775
DKK3	NM_015881	27122	Dickkopf homolog 3 (*Xenopus laevis*)	0.003874938	2.735
CAMKK2	NM_172215	10645	Calcium/calmodulin-dependent protein kinase kinase 2, beta	0.000346607	2.689
C2orf30	NM_015701	27248	Chromosome 2 open reading frame 30	0.003991442	2.652
CDC25A	NM_001789	993	Cell division cycle 25 homolog A (*S. pombe*)	0.000536286	2.634
OXSR1	NM_005109	9943	Oxidative-stress responsive 1	2.57832E-06	2.600
RIN3	NM_024832	79890	Ras and Rab interactor 3	0.000346183	2.582
EEA1	NM_003566	8411	Early endosome antigen 1	3.31698E-05	2.567
HLA-DRB5	NM_002125	3127	Major histocompatibility complex, class II, DR beta 5	0.001474265	2.548
GOT1	AL581249	2805	Glutamic-oxaloacetic transaminase 1, soluble (aspartate aminotransferase 1)	0.003025096	2.529
AGRN	NM_198576	375790	Agrin	0.004551986	2.502
PPM1D	NM_003620	8493	Protein phosphatase 1D magnesium-dependent, delta isoform	0.001089659	2.481
ZBED1	NM_004729	9189	Zinc finger, BED-type containing 1	0.001104421	2.480
ZFAND5	NM_006007	7763	Zinc finger, AN1-type domain 5	0.000927518	2.465
ETS2	NM_005239	2114	V-ets erythroblastosis virus E26 oncogene homolog 2 (avian)	0.000142473	2.459
PTPMT1	BC020242	114971	Protein tyrosine phosphatase, mitochondrial 1	5.88749E-06	2.421
ARHGAP1	NM_004308	392	Rho GTPase activating protein 1	0.003940737	2.403
PRICKLE3	NM_006150	4007	Prickle homolog 3 (*Drosophila*)	0.000505149	2.371
SH3BP4	NM_014521	23677	SH3-domain binding protein 4	0.002583403	2.353
WIPF2	NM_133264	147179	WAS/WASL interacting protein family, member 2	0.003588251	2.317
OTUD1	AB188491	220213	OTU domain containing 1	0.001538639	2.299
DGKD	NM_152879	8527	Diacylglycerol kinase, delta 130 kDa	0.000501851	2.297
RNF31	NM_017999	55072	Ring finger protein 31	0.000146202	2.289
NSFL1C	NM_182483	55968	NSFL1 (p97) cofactor (p47)	0.002998104	2.288
AZIN1	NM_015878	51582	Antizyme inhibitor 1	0.002484085	2.264
ASXL1	NM_015338	171023	Additional sex combs like 1 (*Drosophila*)	0.001918932	2.255
LRRC8A	NM_019594	56262	Leucine rich repeat containing 8 family, member A	0.000265201	2.251
MANBA	NM_005908	4126	Mannosidase, beta A, lysosomal	0.001475313	2.245
PCBD1	NM_000281	5092	Pterin-4 alpha-carbinolamine dehydratase/dimerization cofactor of hepatocyte nuclear factor 1 alpha	0.001161179	2.242
GNB1L	NM_053004	54584	Guanine nucleotide binding protein (G protein), beta polypeptide 1-like	0.000379775	2.241
NMRAL1	NM_020677	57407	NmrA-like family domain containing 1	0.000664578	2.239
EIF1AY	NM_004681	9086	Eukaryotic translation initiation factor 1A, Y-linked	0.001956033	2.234
ASAP1	NM_018482	50807	ArfGAP with SH3 domain, ankyrin repeat and PH domain 1	0.000730755	2.231
C18orf10	NM_015476	25941	Chromosome 18 open reading frame 10	0.000138316	2.228
ZNF330	NM_014487	27309	Zinc finger protein 330	0.00275585	2.216
ATP5G3	NM_001002258	518	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit C3 (subunit 9)	0.000183274	2.211
OGDH	NM_002541	4967	Oxoglutarate (alpha-ketoglutarate) dehydrogenase (lipoamide)	0.000577585	2.211
PIGY	NM_001042616	84992	Phosphatidylinositol glycan anchor biosynthesis, class Y	0.000395444	2.180
MLL5	NM_182931	55904	Myeloid/lymphoid or mixed-lineage leukemia 5 (trithorax homolog, *Drosophila*)	0.001801175	2.172

PAIP1	NM_006451	10605	Poly(A) binding protein interacting protein 1	0.001278086	2.168
SCYL2	NM_017988	55681	SCY1-like 2 (*S. cerevisiae*)	3.57747E-05	2.160
MCTS1	AK096956	28985	Malignant T cell amplified sequence 1	0.000323927	2.137
TRNAU1AP	NM_017846	54952	tRNA selenocysteine 1 associated protein 1	0.001160239	2.137
MAP1A	NM_002373	4130	Microtubule-associated protein 1A	0.001014758	2.131
TNIP1	NM_006058	10318	TNFAIP3 interacting protein 1	0.000254844	2.130
PPP3CC	NM_005605	5533	Protein phosphatase 3 (formerly 2B), catalytic subunit, gamma isoform	0.0004302	2.111
CDC45L	NM_003504	8318	CDC45 cell division cycle 45-like (*S. cerevisiae*)	0.004108453	2.107
KIAA0841	AB020648	23354	KIAA0841	0.00307293	2.095
MTERFD1	NM_015942	51001	MTERF domain containing 1	0.002752437	2.094
INTS12	NM_020395	57117	Integrator complex subunit 12	0.001394525	2.081
EZH2	NM_004456	2146	Enhancer of zeste homolog 2 (*Drosophila*)	0.000585999	2.078
AGPAT3	NM_020132	56894	1-Acylglycerol-3-phosphate O-acyltransferase 3	0.003953771	2.077
PDE6B	NM_000283	5158	Phosphodiesterase 6B, cGMP-specific, rod, beta	0.002411658	2.061
POMGNT1	NM_017739	55624	Protein O-linked mannose beta1,2-N-acetylglucosaminyltransferase	0.001779407	2.054
ACOX1	NM_004035	51	Acyl-coenzyme A oxidase 1, palmitoyl	0.002677004	2.054
BCORL1	AK021694	63035	BCL6 co-repressor-like 1	0.000131308	2.042
ARMC8	AL096748	25852	Armadillo repeat containing 8	0.003651856	2.033
ARMC1	NM_018120	55156	Armadillo repeat containing 1	0.001292391	2.021
SPOCK2	NM_014767	9806	Sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 2	0.003554045	2.017
Overexpressed in failed UBCs					
SSBP3	NM_001009955	23648	Single stranded DNA binding protein 3	0.003603576	4.842
ALKBH6	NM_198867	84964	alkB, alkylation repair homolog 6 (*E. coli*)	0.001413295	3.318
DHX57	NM_198963	90957	DEAH (Asp-Glu-Ala-Asp/His) box polypeptide 57	0.00170115	2.968
NIP30	NM_024946	80011	NEFA-interacting nuclear protein NIP30	0.000183566	2.725
TMEM134	NM_025124	80194	Transmembrane protein 134	0.003386533	2.516
RNPC3	AK057799	55599	RNA-binding region (RNP1, RRM) containing 3	5.00896E-05	2.505
ELK4	NM_001973	2005	ELK4, ETS-domain protein (SRF accessory protein 1)	0.001518504	2.471
TMEM216	NM_016499	51259	Transmembrane protein 216	0.000180326	2.432
GPR150	NM_199243	285601	G protein-coupled receptor 150	0.000487991	2.370
ZFP112	NM_013380	7771	Zinc finger protein 112 homolog (mouse)	4.54855E-05	2.359
ATM	NM_000051	472	Ataxia telangiectasia mutated	0.000542207	2.300
BCAT1	NM_005504	586	Branched chain aminotransferase 1, cytosolic	0.000193566	2.289
PHF2	NM_005392	5253	PHD finger protein 2	0.000820897	2.227
IRX3	NM_024336	79191	Iroquois homeobox 3	0.002209041	2.220
ZNF490	NM_020714	57474	Zinc finger protein 490	0.00040782	2.153
LOC442211	XR_019545	442211	Similar to Vacuolar ATP synthase 16 kDa proteolipid subunit	0.003343215	2.147
PECAM1	NM_000442	5175	Platelet/endothelial cell adhesion molecule	0.000314337	2.131
STON2	NM_033104	85439	Stonin 2	0.000675103	2.096
GSTM2	NM_000848	2946	Glutathione S-transferase mu 2 (muscle)	0.001806696	2.049
GTF2H1	NM_005316	2965	General transcription factor IIH, polypeptide 1, 62 kDa	0.003923314	2.034
DCLRE1A	NM_014881	9937	DNA cross-link repair 1A (PSO2 homolog, *S. cerevisiae*)	0.001829853	2.032
AGAP1	NM_001037131	116987	ArfGAP with GTPase domain, ankyrin repeat and PH domain 1	0.000611158	2.027
HCN4	NM_005477	10021	Hyperpolarization activated cyclic nucleotide-gated potassium channel 4	0.001048363	2.002

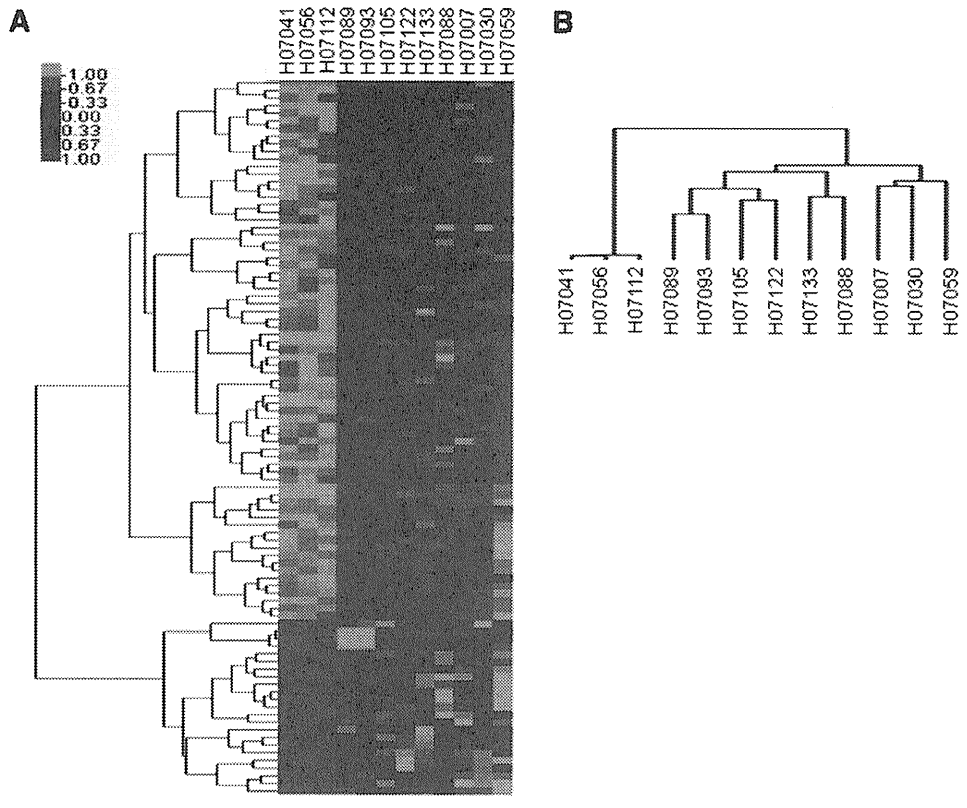

Fig. 3. Differentially expressed genes in successful and failed UCB samples. (A) Gene expression heat map of human UCB samples showing differential expression in successful and failed UCB samples. The map shows genes whose expression showed a larger than 2-fold increase or decrease and were statistically significant ($p < 0.005$) (see Table 1). Genes that were expressed at above or below the average level of the twelve samples are indicated in red and green, respectively. The color bar is the logarithmic indicator of the fold difference of each gene expression from the average, i.e., 1.0 means a larger than 2-fold increase and −1.0 means a less than 0.5-fold decrease. (B) Dendrogram of human UCB gene expression profiles constructed using the differentially expressed genes.

Initially, we employed a hierarchical clustering analysis to profile gene expression in the human CD34$^+$ UCB cells. However, this analysis did not separate successful and failed UCBs (data not shown). We concluded that our collection of UCBs might not be large enough to extract biologically meaningful signatures from gene expression profiles through an unsupervised approach.

In order to identify candidate genes responsible for success or failure of engraftment, we compiled a list of genes showing differential expression between successful and failed UCBs. A combination of relative level (fold change) of expression and statistical significance (Student's *t*-test) was used to distinguish these genes. Genes that showed a larger than 2-fold increase or decrease in gene expression between successful and failed UCBs, and also showed a statistically significant difference ($p < 0.005$) are listed in Table 1.

In total, 71 genes were found to be upregulated in successful UCBs and 23 in failed UCBs (Table 1). Many of the genes showing upregulation in successful UCBs are important for cell growth and differentiation in hematopoietic cells, such as *HOXB4*, *ETS2*, *CDC45L*, and *SMC1A* (Table 1).

In the expression heat map for the genes listed in Table 1, genes that expressed above or below the average level of the twelve UCB samples are indicated in red or green, respectively (Fig. 3A). The dendrogram was obtained by cluster analyses based on the differentially expressed genes (Fig. 3B). The graph indicates that failed UCBs were clearly separate from successful UCBs and consisted of a single cluster (Fig. 3B).

3.3. GSEA and DAVID analyses confirm upregulation of cell growth related genes in successful UCBs

Following identification of differentially expressed genes, we performed GSEA analyses in order to obtain biologically relevant insights. Four gene sets were selected as specifically enriched in

Table 2
Gene sets significantly overrepresented in successful and failed UCBs.

Gene set name	# of genes	ES	NES	NOM p-val	FDR q-val	FWER p-val
Enriched in successful UCBs						
SCHUMACHER MYC UP	53	0.561	2.085	0.000	0.026	0.021
ZHAN MM CD138 PR VS REST	23	0.648	2.007	0.000	0.048	0.077
P21_ANY_DN	22	0.611	1.859	0.000	0.193	0.473
UVB_NHEK2_DN	74	0.477	1.869	0.000	0.228	0.442
Enriched in failed UCBs						
DAC_PANC50_UP	26	−0.597	−1.976	0.000	0.127	0.167

successful UCBs (Table 2). Three of the four sets of genes have an unambiguous role in cell growth activity: the SCHUMACHER_MYC_UP gene set consists of downstream genes of the *MYC* oncogene in B-cells [20]; the ZHAN_MM_CD138_PR_VS_REST gene set is overexpressed in multiple myelomas with poor prognosis [21]; the P21_ANY_DN gene set includes genes that are downregulated by the tumor suppressor p21 [22]. However, it is currently not clear whether the UVB_NHEK2_DN gene set is active in cell growth [23]. The results of the GSEA analysis suggest that successful UCBs may be more committed to cell growth than the failed UCBs. All gene sets that were significantly enriched in either successful or failed UCBs are listed in *Supplementary* Table 3.

The DAVID annotation service was also employed for the functional analysis of differentially regulated genes in successful and failed UCBs. For this analysis, we chose a 1.5-fold increase or decrease in gene expression with $p < 0.005$ as the criteria for gene selection from successful and failed UCBs expression profiles. These criteria resulted in the selection of 577 and 327 genes, respectively, from the gene expression profiles of successful and failed UCBs (*Supplementary* Table 4). The GO terms for the biological processes of the significantly enriched genes (fold enrichment >1.5, FDR <5%) for both groups are given in Table 3. From examination of Table 3, it is clear that the successful UCBs had high expression of cell cycle related genes, such as those in "GO:0007049 cell cycle", whereas developmental and morphogenesis-related genes were upregulated in failed UCBs.

3.4. HOXB4 and other cell cycle related genes are upregulated in successful UCBs

Next, we sought to confirm the conclusion from the microarray analysis that cell growth related genes were overexpressed in successful UCBs. We performed real-time RT-PCR analyses (qPCR) of four genes of interest: *CDC45L*, *C/EBP-β*, *ETS2*, and *HOXB4*. Due to the limited amount of cDNA available, only samples from six UCB samples (H07007, H07041, H07056, H07089, H07093, and H07015) could be used for qPCR. Of these, H07041 and H07056 are failed UCBs. We found that the amount of normalized qPCR product from the four mRNAs showed a good correlation with the signal intensity of the corresponding microarray probes (Fig. 4).

4. Discussion

The present study indicated that $CD34^+$ cells from 12 different human UCBs showed various abilities to reconstitute hematopoiesis in sublethally irradiated NOD/SCID mice. Gene expression profiling of these UCBs suggested that those that were successful at engraftment had increased expression of genes associated with cell growth compared to failed UCBs. To date, this is the first report to describe a relationship between the engraftment ability of mouse xenografts and gene expression profiles in human $CD34^+$ UCB cells. Indeed, our results suggest that the gene expression profile of human $CD34^+$ UCB cells reflect their potential for successful establishment of hematopoiesis in UCB transplantation.

There are several reports describing the gene expression profiles of human UCBs [24–27]. However, these studies provided no information on the relationship between gene expression profiles in human $CD34^+$ UCB cells and their relative abilities for bone marrow engraftment in mouse xenografts. We found that different cryopreserved human $CD34^+$ UCB cells varied in the extent of engraftment they yielded in mouse xenografts. This variation raises the question of what factors determine successful engraftment by cryopreserved human $CD34^+$ UCB cells? It is well known that the relative numbers of hematopoietic stem cells is a critical quality factor for UCBs [5,8,10]. The study by Cairo et al. further showed that the colony formation activity of human UCBs was correlated with ethnicity, sex, and the delivery methods of the donors [11]. It is an open question as to why these factors should be correlated with the number of stem cells in UCBs.

Several transcription factors are candidate mediators of cell growth for human $CD34^+$ UCB cells, for example, *HOXB4* and *ETS2*, which were upregulated more than 2-fold in successful UCBs compared to failed UCBs (Table 1). *HOXB4* is a major factor for the growth and maintenance of 'stemness' in embryonic stem cells [28]. Several groups have reported that introduction of *HOXB4* into UCB cells contributed to the *ex vivo* expansion of cell numbers [29,30]. *ETS2* is an oncogene that plays critical roles in cell growth signal transduction in various tissues [31–33]. Our results here are compatible with the known characteristics of these transcription factors.

C/EBP-β is another transcription factor upregulated in successful UCBs. The *C/EBP-β* is a critical factor for cell differentiation and expansion of the number of progenitor cells committed to the B-cell lineage; it also promotes tumor growth in several types of malignancies [34]. In contrast, the tumor suppressor *ATM* was included in the set of upregulated genes in the failed UCBs. *ATM* is activated by DNA damage and can induce cell cycle arrest [35].

Our experimental approach demonstrates the practicality of molecular assessment of the quality of human $CD34^+$ UCB cells. At present, it is not clear whether any candidate cell surface

Table 3
GO terms for biological processes enriched in successful and failed UCBs (>1.5-fold, FDR <5%).

GO category	Term	Count	%	*p*-Value	Fold enrichment	FDR (%)
Enriched in successful UCBs						
GO:0006512	Ubiquitin cycle	38	6.60	0.00	2.51	0.00
GO:0043687	Post-translational protein modification	71	12.33	0.00	1.59	0.19
GO:0006888	ER to Golgi vesicle-mediated transport	9	1.56	0.00	4.78	1.03
GO:0006281	DNA repair	19	3.30	0.00	2.42	1.80
GO:0007049	Cell cycle	44	7.64	0.00	1.67	1.86
GO:0046907	Intracellular transport	37	6.42	0.00	1.76	2.13
GO:0065003	Macromolecular complex assembly	31	5.38	0.00	1.88	2.17
GO:0015031	Protein transport	36	6.25	0.00	1.76	2.70
GO:0048193	Golgi vesicle transport	11	1.91	0.00	3.40	2.78
GO:0022607	Cellular component assembly	32	5.56	0.00	1.81	3.37
GO:0000074	Regulation of progression through cell cycle	29	5.03	0.00	1.87	3.43
GO:0051726	Regulation of cell cycle	29	5.03	0.00	1.86	3.71
Enriched in failed UCBs						
GO:0019222	Regulation of metabolic process	60	20.34	0.00	1.54	0.66
GO:0009653	Anatomical structure morphogenesis	29	9.83	0.00	1.96	1.47
GO:0031323	Regulation of cellular metabolic process	57	19.32	0.00	1.52	1.53
GO:0050793	Regulation of developmental process	11	3.73	0.00	3.25	4.04

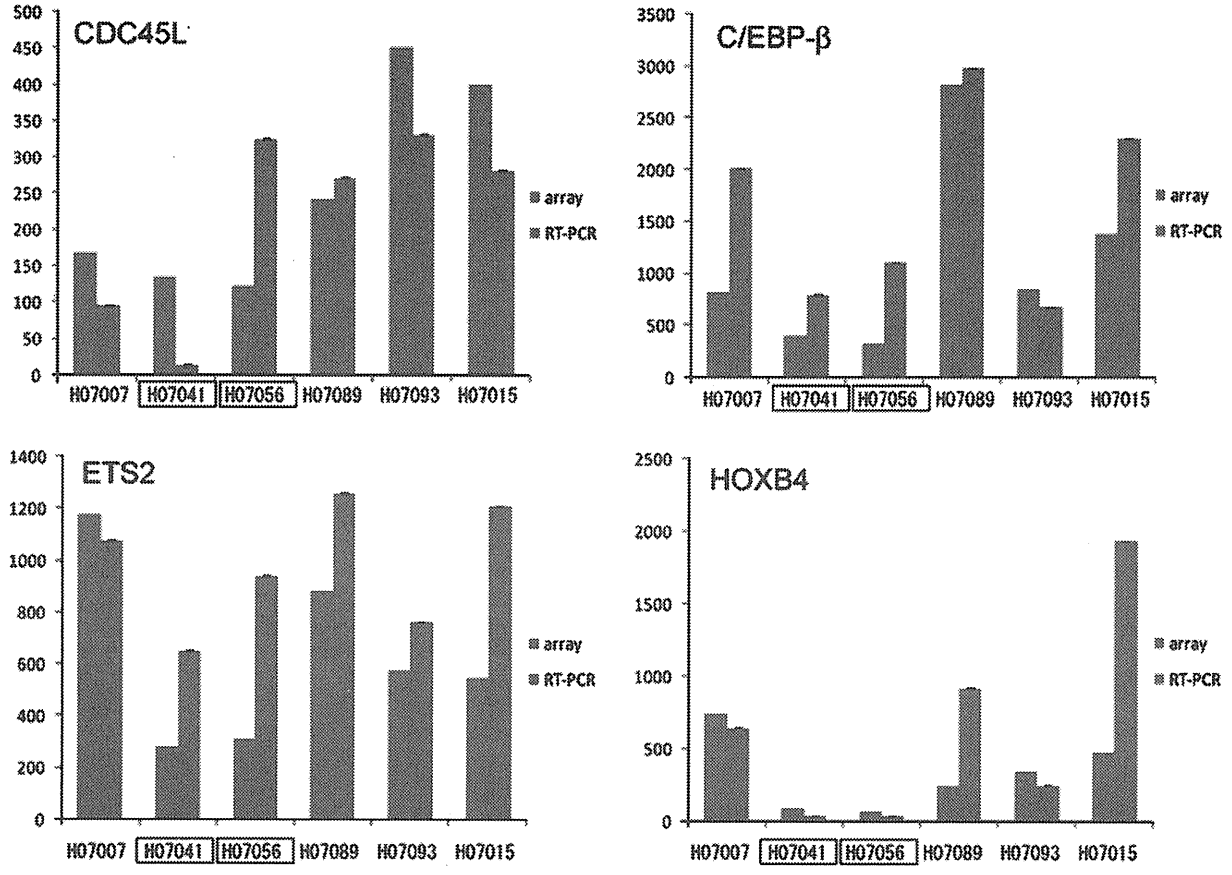

Fig. 4. A quantitative RT-PCR analysis was used to evaluate overexpressed genes in the successful UCB samples: the results for *CDC45L*, *C/EBP-β*, *ETS2*, and *HOXB4* are shown. The vertical axis indicates the relative expression ratio of each gene normalized against GAPDH. The names under the horizontal axis indicate the different human UCB samples and the boxed names indicate those that failed to engraft (see text).

markers for successful UCBs are included among the identified upregulated genes (Table 2). Nevertheless, it will be valuable to establish robust molecular markers for potentially successful $CD34^+$ UCB cells using functional gene expression profiling.

5. Conclusions

The quality of cryopreserved human $CD34^+$ UCB cells was variable and their respective gene expression profiles might reflect these qualitative differences and provide clinically relevant and versatile surrogate markers for human $CD34^+$ UCB cell quality. In addition, the results in this study suggest that cell growth is an important trait for the successful engraftment of human $CD34^+$ UCB cells.

Support and financial disclosure declaration

This work was supported by a grant from the Ministry of Education, Culture, Sports, Science, and Technology in Japan (MEXT) through the RIKEN Strategic Research Programs for Research and Development and was also supported by a Grant-in-aid for Scientific Research to J.Y. and by Global COE Program (Network Medicine), MEXT. All authors declare no competing financial interests.

Acknowledgments

We thank Drs. Tetsuo Noda and Yoshihide Hayashizaki for providing resources, and the DNA Chip Research Inc. (Yokohama, Japan) for its technical assistance. We obtained human UCBs from the Cell Engineering Division of RIKEN BioResource Center, which is supported by the Project for Realization of Regenerative Medicine and the National Bio-Resources Project of the MEXT.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2010.06.010.

References

[1] V.K. Prasad, J. Kurtzberg, Umbilical cord blood transplantation for non-malignant diseases, Bone Marrow Transplant. 44 (2009) 643–651.
[2] V. Rocha, M. Labopin, G. Sanz, et al., Transplants of umbilical-cord blood or bone marrow from unrelated donors in adults with acute leukemia, N. Engl. J. Med. 351 (2004) 2276–2285.
[3] J.A. Brown, V.A. Boussiotis, Umbilical cord blood transplantation: basic biology and clinical challenges to immune reconstitution, Clin. Immunol. 127 (2008) 286–297.
[4] B.G. Thomson, K.A. Robertson, D. Gowan, et al., Analysis of engraftment, graft-versus-host disease, and immune recovery following unrelated donor cord blood transplantation, Blood 96 (2000) 2703–2711.
[5] C.A. Rodrigues, G. Sanz, C.G. Brunstein, et al., Analysis of risk factors for outcomes after unrelated cord blood transplantation in adults with lymphoid malignancies: a study by the Eurocord–Netcord and lymphoma working party of the European group for blood and marrow transplantation, J. Clin. Oncol. 27 (2009) 256–263.
[6] M.J. Laughlin, J. Barker, B. Bambach, et al., Hematopoietic engraftment and survival in adult recipients of umbilical-cord blood from unrelated donors, N. Engl. J. Med. 344 (2001) 1815–1822.
[7] P. Rubinstein, C. Carrier, A. Scaradavou, et al., Outcomes among 562 recipients of placental-blood transplants from unrelated donors, N. Engl. J. Med. 339 (1998) 1565–1577.
[8] J.E. Wagner, J.N. Barker, T.E. DeFor, et al., Transplantation of unrelated donor umbilical cord blood in 102 patients with malignant and nonmalignant

diseases: influence of CD34 cell dose and HLA disparity on treatment-related mortality and survival, Blood 100 (2002) 1611–1618.
[9] E. Gluckman, V. Rocha, W. Arcese, et al., Factors associated with outcomes of unrelated cord blood transplant: guidelines for donor choice, Exp. Hematol. 32 (2004) 397–407.
[10] S.S. Grewal, J.N. Barker, S.M. Davies, et al., Unrelated donor hematopoietic cell transplantation: marrow or umbilical cord blood? Blood 101 (2003) 4233–4244.
[11] M.S. Cairo, E.L. Wagner, J. Fraser, et al., Characterization of banked umbilical cord blood hematopoietic progenitor cells and lymphocyte subsets and correlation with ethnicity, birth weight, sex, and type of delivery: a Cord Blood Transplantation (COBLT) Study report, Transfusion 45 (2005) 856–866.
[12] C.E. Stevens, J. Gladstone, P.E. Taylor, et al., Placental/umbilical cord blood for unrelated-donor bone marrow reconstitution: relevance of nucleated red blood cells, Blood 100 (2002) 2662–2664.
[13] O.I. Gan, B. Murdoch, A. Larochelle, et al., Differential maintenance of primitive human SCID-repopulating cells, clonogenic progenitors, and long-term culture-initiating cells after incubation on human bone marrow stromal cells, Blood 90 (1997) 641–650.
[14] J.E. Dick, M. Bhatia, O. Gan, et al., Assay of human stem cells by repopulation of NOD/SCID mice, Stem cells 15 (Suppl. 1) (1997) 199–203; discussion, 204–197
[15] A. Dafforn, P. Chen, G. Deng, et al., Linear mRNA amplification from as little as 5 ng total RNA for global gene expression analysis, Biotechniques 37 (2004) 854–857.
[16] M.J. de Hoon, S. Imoto, J. Nolan, et al., Open source clustering software, Bioinformatics 20 (2004) 1453–1454.
[17] A.J. Saldanha, Java Treeview – extensible visualization of microarray data, Bioinformatics 20 (2004) 3246–3248.
[18] A. Subramanian, P. Tamayo, V.K. Mootha, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. USA 102 (2005) 15545–15550.
[19] W. Huang da, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. Protoc. 4 (2009) 44–57.
[20] M. Schuhmacher, F. Kohlhuber, M. Holzel, et al., The transcriptional program of a human B cell line in response to Myc, Nucleic Acids Res. 29 (2001) 397–406.
[21] F. Zhan, Y. Huang, S. Colla, et al., The molecular classification of multiple myeloma, Blood 108 (2006) 2020–2028.
[22] Q. Wu, P. Kirschmeier, T. Hockenberry, et al., Transcriptional regulation during p21WAF1/CIP1-induced apoptosis in human ovarian cancer cells, J. Biol. Chem. 277 (2002) 36329–36337.
[23] J. Takao, K. Ariizumi, I. Dougherty, et al., Genomic scale analysis of the human keratinocyte response to broad-band ultraviolet-B irradiation, Photodermatol. Photoimmunol. Photomed. 18 (2002) 5–13.
[24] M. Merkerova, A. Vasikova, H. Bruchova, et al., Differential gene expression in umbilical cord blood and maternal peripheral blood, Eur. J. Haematol. 83 (2009) 183–190.
[25] J.A. Jeong, S.H. Hong, E.J. Gang, et al., Differential gene expression profiling of human umbilical cord blood-derived mesenchymal stem cells by DNA microarray, Stem Cells 23 (2005) 584–593.
[26] E. Martin-Rendon, S.J. Hale, D. Ryan, et al., Transcriptional profiling of human cord blood CD133$^+$ and cultured bone marrow mesenchymal stem cells in response to hypoxia, Stem Cells 25 (2007) 1003–1012.
[27] Y.Y. Ng, B. van Kessel, H.M. Lokhorst, et al., Gene-expression profiling of CD34$^+$ cells from various hematopoietic stem-cell sources reveals functional differences in stem-cell activity, J. Leukoc. Biol. 75 (2004) 314–323.
[28] B.P. Sorrentino, Clinical strategies for expansion of haematopoietic stem cells, Nat. Rev. Immunol. 4 (2004) 878–888.
[29] J. Antonchuk, G. Sauvageau, R.K. Humphries, HOXB4-induced expansion of adult hematopoietic stem cells ex vivo, Cell 109 (2002) 39–45.
[30] J. Krosl, P. Austin, N. Beslu, et al., In vitro expansion of hematopoietic stem cells by recombinant TAT-HOXB4 protein, Nat. Med. 9 (2003) 1428–1432.
[31] T. Hsu, M. Trojanowska, D.K. Watson, Ets proteins in biological control and cancer, J. Cell Biochem. 91 (2004) 896–903.
[32] J.S. Yordy, R.C. Muise-Helmericks, Signal transduction and the Ets family of transcription factors, Oncogene 19 (2000) 6503–6513.
[33] G. Mavrothalassitis, J. Ghysdael, Proteins of the ETS family with transcriptional repressor activity, Oncogene 19 (2000) 6524–6532.
[34] C. Nerlov, The C/EBP family of transcription factors: a paradigm for interaction between gene expression and proliferation control, Trends Cell Biol. 17 (2007) 318–324.
[35] Y. Xu, D. Baltimore, Dual roles of ATM in the cellular response to radiation and in cell growth control, Genes Dev. 10 (1996) 2401–2410.

UICC
global cancer control

IJC
International Journal of Cancer

Check your cultures! A list of cross-contaminated or misidentified cell lines

Amanda Capes-Davis[1], George Theodosopoulos[1], Isobel Atkin[2], Hans G. Drexler[3], Arihiro Kohara[4], Roderick A.F. MacLeod[3], John R. Masters[5], Yukio Nakamura[6], Yvonne A. Reid[7], Roger R. Reddel[1] and R. Ian Freshney[8]

[1] CellBank Australia – Children's Medical Research Institute, Westmead, NSW, Australia
[2] European Collection of Cell Cultures (ECACC) – Health Protection Agency, Porton Down, Salisbury, Wiltshire, United Kingdom
[3] DSMZ – German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany
[4] JCRB – Japanese Collection of Research Bioresources, Osaka, Japan
[5] Institute of Urology, University College London, London, United Kingdom
[6] RIKEN – BioResource Center Cell Engineering Division, Tsukuba, Japan
[7] ATCC – American Type Culture Collections, Manassas, VA
[8] Centre for Oncology and Applied Pharmacology, Glasgow University, Glasgow, United Kingdom

Mini Review

Continuous cell lines consist of cultured cells derived from a specific donor and tissue of origin that have acquired the ability to proliferate indefinitely. These cell lines are well-recognized models for the study of health and disease, particularly for cancer. However, there are cautions to be aware of when using continuous cell lines, including the possibility of contamination, in which a foreign cell line or microorganism is introduced without the handler's knowledge. Cross-contamination, in which the contaminant is another cell line, was first recognized in the 1950s but, disturbingly, remains a serious issue today. Many cell lines become cross-contaminated early, so that subsequent experimental work has been performed only on the contaminant, masquerading under a different name. What can be done in response—how can a researcher know if their own cell lines are cross-contaminated? Two practical responses are suggested here. First, it is important to check the literature, looking for previous work on cross-contamination. Some reports may be difficult to find and to make these more accessible, we have compiled a list of known cross-contaminated cell lines. The list currently contains 360 cell lines, drawn from 68 references. Most contaminants arise within the same species, with HeLa still the most frequently encountered (29%, 106/360) among human cell lines, but interspecies contaminants account for a small but substantial minority of cases (9%, 33/360). Second, even if there are no previous publications on cross-contamination for that cell line, it is essential to check the sample itself by performing authentication testing.

Key words: authentication, cell culture, cell lines, cross-contamination, DNA profiling, misidentification
Additional Supporting Information may be found in the online version of this article.
Novelty and Impact: This manuscript reviews the literature relating to cross-contamination of cell lines. Its novelty comes from the inclusion of a list of known cross-contaminated cell lines (over 300 lines named), allowing researchers to check their own cell lines with reference to the article. Recent developments in this field, including methods of authentication testing, are also discussed.
Grant sponsor: National Health and Medical Research Council of Australia
DOI: 10.1002/ijc.25242
History: Received 24 Nov 2009; Accepted 18 Jan 2010; Online 8 Feb 2010
Correspondence to: Amanda Capes-Davis, CellBank Australia, Children's Medical Research Institute, Locked Bag 23, Wentworthville, NSW 2145, Australia, Fax: +61 2 9687 2120, E-mail: acapdav@gmail.com

Cell Lines as Model Systems

Continuous cell lines represent a readily accessible and easily studied resource for research into health and disease. These cell lines have acquired the ability to proliferate indefinitely if grown in the appropriate culture conditions; usually this is a rare event, since the majority of cells even in tumor tissue will cease proliferation after a limited number of cell divisions.[1] However, once established, a continuous cell line can be repeatedly passaged, reliably recovers from cryopreservation and retains many of the properties of its cell type or tissue of origin.[2,3] These advantages make continuous cell lines effective, and widely used, model systems for normal cellular processes and for a variety of disease states.

Cell lines are particularly attractive models for studying malignant disease. The genetic changes in tumor-derived cell lines closely resemble those of the tumors of origin.[4] Moreover, the genetic changes required to establish continuous cell lines from normal cells recapitulate many of the genetic changes occurring in cancer.[5,6] These genetic changes are required to overcome replicative senescence, in which normal cells continue to be metabolically active but are restricted from further division.[1] Cells able to overcome senescence continue

proliferating until their telomeres become so short that the chromosomes undergo fusion-breakage-bridge cycles and the ensuing genomic instability results in culture crisis. Occasionally (at a rate of ~1 in 10^7 cells), an immortalized cell will emerge from crisis and begin to divide again, yielding a continuous cell line.[1] The changes seen throughout this process have many parallels within cancer development, both for malignancy in general and when considering specific tumor types.[7,8]

Despite these advantages, numerous cautions have emerged from the literature regarding appropriate use of cell lines as model systems.[9,10] Even where cultures have been transformed through the introduction of specific genes, cell lines that have passed through replicative senescence and crisis are aneuploid, heteroploid and genotypically and phenotypically unstable, resulting in considerable heterogeneity within the culture.[10] This instability will cause changes in the characteristics of the cell line but a further consequence may result: alterations in a cell line can be accepted by the user as intrinsic to that culture when there is actually extrinsic contamination present.

Cell Line Cross-contamination and Misidentification

Cell lines become contaminated when a foreign cell line or microorganism is introduced without the handler's knowledge. Although we do not wish to minimize the problem of microbial contamination, we will focus on cell line cross-contamination in this article. Cross-contamination may arise due to several causes, including poor technique (spread *via* aerosols or accidental contact), use of unplugged pipets, sharing media and reagents among cell lines and use of mitotically inactivated feeder layers or conditioned medium, which may carry contaminating cells if not properly eliminated, for example, by freeze-thaw and filtration.[11] In addition, a cell line can be replaced by another as a result of misidentification by confusing cultures during handling, mislabeling or poor freezer inventory control. Simple errors during labeling of culture flasks, truncation of the cell line name or typographic errors in a published manuscript, can result in significant confusion for years after the event when another researcher attempts to use the same cell line for ongoing experimental work.[12]

Cross-contamination may occur "early," in which case the original cell line has probably never existed independently, or "late," where the tested sample has been overgrown but other stocks of the original may still exist.[13] Unfortunately, cell lines generally become cross-contaminated early, while still within the originating laboratory.[14] This is not surprising: cultures can remain in crisis for a prolonged period of time before emergence of an immortalized population and this is a time when a single cell, if introduced from a separate cell line, would rapidly take over the culture.

There are now a number of studies pointing out the severity of this problem and the need to take urgent action to minimize cross-contamination and its consequences.[9,15–17] Ten years ago, the German Collection of Microorganisms and Cell Cultures (DSMZ) published data from its identification testing of cancer cell lines submitted by various laboratories for deposit at the cell bank.[14] They found that 18% of 252 submitted cell lines were cross-contaminated with more than half of cases arising within only 6 laboratories. Subsequent work by the DSMZ, extending the number of cell lines tested (Fig. 1), shows that of 598 leukemia–lymphoma cell lines (the group provided with the most complete genetic data), 187 (31%) were contaminated with Mycoplasma and/or a second cell line with 38 (6%) of cell lines contaminated with both. These data suggest that poor practice within some laboratories results in contamination of multiple cell lines with multiple contaminants, which can then be disseminated more widely if these cultures are used by others.

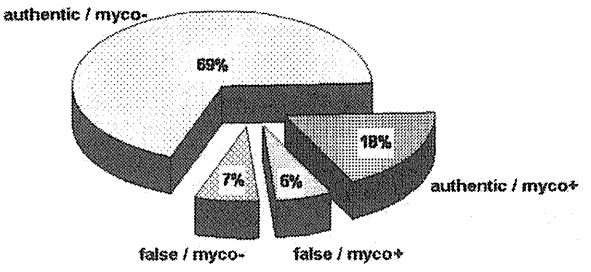

Figure 1. Rates of contamination for leukemia–lymphoma cell lines. Percentages of cross-contaminated and Mycoplasma-contaminated cell lines from a dataset of 598 leukemia and lymphoma cell lines analyzed by the German cell line bank DSMZ. "False/authentic" refers to the presence or absence of cross-contamination; "myco+/myco−" refers to the presence or absence of Mycoplasma contamination. Cell lines fall into the following categories: authentic/myco− (n = 411, 69%); authentic/myco+ (n = 108, 18%); false/myco− (n = 41, 7%) and false/myco+ (n = 38, 6%). (Courtesy of Hans Drexler, DSMZ.)

Other studies have pointed out that testing of cell lines is often infrequent, resulting in the failure to detect contaminated samples. John Ryan of Corning Life Sciences conducted surveys of seminar attendees in 1990, asking about Mycoplasma contamination; 50% were not currently performing testing and only 18% said they tested their cultures regularly. Almost 1 in 4 respondents (23%) had experienced Mycoplasma contamination, but with such a low level of testing, it is likely that the real figure was much higher.[18] Other data on cross-contamination were published in 2004 by researchers at the University of California, Berkeley, where Walter Nelson-Rees worked on this problem in the 1970s, focusing on the HeLa cell line.[19] Of 483 respondents to a questionnaire on cell line usage, 35% were using cell lines obtained from another laboratory rather than a cell line repository, but almost half of all respondents performed no testing for cross-contamination.[20]

A practical example of the consequences of cell line contamination can be found in a recent study published by Berglind *et al.*[21] The authors analyzed data within the UMD_p53 (2007) database, which includes information on the p53 status of 1,211 cell lines. Discrepancies were found in p53 status for 23% (88/384) of cell lines where data have been published by 2

independent laboratories. It is likely that many of these discrepancies arose due to work with cross-contaminated samples; the authors noted that many groups rely on previously published reports of a cell line's p53 status,[21] resulting in further confusion when interpreting results from these cell lines.

Cell banks have the expertise to detect such cross-contamination, and have been proactive in publishing reports of cross-contaminated cell lines,[22,23] in publishing test results online[24] and in developing new detection methods.[25–27] Unfortunately, however, cell banks have also reported reluctance from many researchers to deposit cell lines for distribution.[28] Such repositories specialize in the detection of cross-contamination and it is unlikely that most laboratories have comparable resources in this regard. In addition, many researchers obtain cell lines from one another, rather than approaching the originator or purchasing the cell line from a cell bank performing quality control testing. This may be faster or cheaper than obtaining cultures from a reputable source but the practice makes contamination more prevalent and harder to detect.

Practical Responses

Having defined the problems, it is time to focus on what can be done. Several cancer-related journals, including the International Journal of Cancer, have recently responded to these issues by changing their policies to require evidence of authentication with all submitted manuscripts using continuous cell lines.[29,30] Their response underscores the need for laboratories to come to grips with cell line cross-contamination and misidentification. Every researcher involved in cell culture will have cell lines currently in culture, stored in liquid nitrogen or may be commencing work on a new cell line. Put practically, how can you know if your cell lines are cross-contaminated?

There are 2 important answers to this question:

1. Check the literature, for example, by searching the PubMed database using the cell line name and "cross-contamination."
2. Check your cultured cells. Unless a cell line has come directly from a repository or other laboratory performing identification testing, it should be tested on arrival, and all cultures should be periodically tested while in use, before cryopreservation and when thawed from liquid nitrogen.[31] A variety of methods are available for authentication; for human cell lines, short tandem repeat (STR) profiling is the current international reference standard and is recommended as an easy and economical way to confirm cell line identity by comparison to donor tissue or to other samples of the cell line held by laboratories worldwide.[26]

Checking the Literature: A List of Cross-Contaminated Cell Lines

A 2004 survey of abstracts within the PubMed database would suggest that inappropriate usage of cross-contaminated cell lines is increasing,[20] despite many years of publication on this issue. It is possible that many researchers simply cannot find existing references to cross-contamination so, to make this already published work more accessible, we have surveyed the literature and other online resources for references to cell line contamination. The resulting list of cross-contaminated cell lines is included as Electronic Supporting Information.

To generate this list, the authors examined the PubMed database, references within other articles relating to this topic and the websites of 5 cell banks: the American Type Culture Collection (ATCC), DSMZ, European Collection of Cell Cultures (ECACC), Japanese Collection of Research Bioresources and the RIKEN Bioresource Center Cell Bank. A Wikipedia list of contaminated cell lines was also accessed (http://en.wikipedia.org/wiki/List_of_contaminated_cell_lines). Cross-contaminated cell lines are listed by name along with their species and cell type (both claimed and actual), the name of the contaminating cell line where identified, the reference in which this was reported and the PubMed ID number where available. Notes are also included for some cell lines. The list is made available in Excel spreadsheet or PDF format for easy accessibility.

The cell lines listed within this database are divided into 2 tables. Supporting Information Table 1 contains those cell lines where cross-contamination occurred as an early event, and thus where there is no original material remaining. Supporting Information Table 2 contains those cell lines where it is thought cross-contamination occurred as a late event and where original stocks may still exist. A full list of references is also given.

The current list of cross-contaminated cell lines (version 6.4) contains 360 cell lines, 346 in Supporting Information Table 1 and 14 in Supporting Information Table 2, drawn from 68 references. Cell lines affected are primarily human, although cultures from at least 8 other species are included, and come from a wide spectrum of tissue types. The cell or tumor type is given within the list where known; extensive work has been done by some cell banks and laboratories in this area to characterize the actual cell type or tumor type.[22,32] In some cases, this work has shown that a cell line carries the correct name but its cell or tumor type has been incorrectly identified, for example, the cell line RPMI-6666 was initially thought to have come from Hodgkin lymphoma but is now known to be an EBV-positive B-lymphoblastoid cell line.[22]

Common features for cross-contaminating cell lines within the current list are summarized in Table 1. It can be seen that most cross-contamination events have arisen from within the same species but a substantial minority (9%, 33/360) involved cross-contamination from a second species. For the intraspecies contaminants, all of those detected were human but it is likely that this relates to the difficulty of detecting intraspecies contaminants for nonhuman species. The commonest contaminant remains the HeLa cell line

Table 1. Cross-contaminating cell lines

Type of contaminant	Number of cell lines affected
Intraspecies	
Human	324
Nonhuman	0
Interspecies	33
Correct name—incorrect cell type (misidentified)[1]	3
Total	360
Contaminating cell line—12 most frequent	**Number of cell lines affected**
HeLa (human cervical adenocarcinoma)	106
T-24 (human bladder carcinoma)	18
HT-29 (human colon carcinoma)	12
CCRF-CEM (human acute lymphoblastic leukemia)	9
K-562 (human chronic myeloid leukemia)	9
U-937 (human lymphoma)	8
OCI/AML2 (human acute myeloid leukemia)	8
Hcu-10 (human esophageal carcinoma)[2]	7
M14 (human melanoma)	7
HL-60 (human acute myeloid leukemia)	6
PC3 (human prostate carcinoma)	6
SW-480, SW620 (human colon carcinoma)[3]	6

[1]For additional misidentified cell lines see Drexler *et al.*[22] [2]Hcu-10 carries the same genetic identity as Hcu-18, Hcu-22, Hcu-27, Hcu-33, Hcu-37 and Hcu-39; it is unclear which is the correct identity (see Electronic Supporting Information for reference). [3]SW480 and SW620 come from the same donor and therefore carry the same genetic identity (see Electronic Supporting Information for reference).

(29%, 106/360), followed by T-24 (5%, 18/360) and HT-29 (3%, 12/360).

It is important for such a list to be continually updated and feedback is welcome for this purpose. An earlier version of the database was released online by ECACC[31]; 6 cell banks have now agreed to make the database available online and to update this information where necessary. Current website addresses for access to the list of cross-contaminated cell lines are given in Table 2. In future, it is envisaged that the current list of misidentified cell lines will be included in a new initiative improving access to authentication data. The Standard Development Organization at the ATCC is in the process of producing an international standard for human cell line identification based on STR profiling (ATCC SDO Workgroup ASN-0002, manuscript submitted). Strict criteria for STR profiles derived from cancer cell lines are being developed. One consequence of this initiative is that funding is being sought for a quality controlled and curated cell line database with free access into which the database described here will be incorporated.

Table 2. Websites for ongoing access to the list of cross-contaminated cell lines

Cell bank	Website address
ATCC	http://www.atcc.org/
CellBank Australia	http://www.cellbankaustralia.com/
DSMZ	http://www.dsmz.de/
ECACC	http://www.hpacultures.org.uk/collections/ecacc.jsp
JCRB	http://cellbank.nibio.go.jp/
RIKEN Bioresource Center Cell Bank	http://www.brc.riken.go.jp/lab/cell/english/guide.shtml

Checking Your Cultures: Authentication of Cell Lines

Even if a search of the literature shows no indication that a cell line is contaminated, it is still essential to test the sample that you are working with. Authentication testing should be considered in a positive light, as an essential part of good cell culture practice[33] and as an assurance for researchers, funding bodies and journals that the cell line used is a valid experimental model.[17]

There are a number of methods for testing cell line identity. When the issue of cross-contamination was first identified, HeLa contaminants were detected through a combination of isoenzyme and chromosomal analysis.[19,34] Both techniques continue to be used but there are also many newer molecular approaches. Commonly used authentication methods are summarized in Table 3; what factors should be considered when choosing between these methods?

The expertise of the laboratory holding the cell line is an important factor. For example, laboratories with experience in cytogenetics would have the skills to identify species through karyotype analysis and cell lines through the presence or absence of appropriate markers.[35] Although this is an older approach, it still allows clear identification of cell lines, and many cell banks have published karyotypic information on their cell lines to allow comparison to well-characterized stocks. It should be noted that tumor-derived cell lines can be surprisingly difficult to harvest for cytogenetic analysis[35] and are typically heteroploid making interpretation difficult: the experience of the operator is important for success.

The species of cell lines held within the laboratory is also important. Although some authentication methods can be used on more than 1 species, molecular methods such as STR profiling are only successful for a single species; other species will simply fail to amplify.[26] This may not be an issue for laboratories working only with human samples but clearly is a significant factor for groups working with rodent cell lines. In this regard, multilocus DNA fingerprint analysis has a clear advantage, since probes are able to hybridize to a wide variety of species.[25] Unfortunately, although successful within a single laboratory, it can be challenging to compare DNA fingerprints across several experimental runs, and it is difficult to exchange data among laboratories or for cell

Table 3. Commonly used methods for authenticating cell lines

Name	Description	Purpose	References
Chromosomal analysis/karyotyping	Involves preparation of a metaphase spread with chromosome banding and painting to identify chromosome number and markers	Separates species, plus individual cell lines if detailed analysis performed	Ref. 35
Isoenzyme analysis	Biochemical method separating isoenzymes by electrophoresis; isoenzyme mobility may vary within or across species. Kits available include the Authentikit gel electrophoresis system	Separates species, sometimes individuals	Refs. 36,37
Multilocus DNA fingerprint analysis	Molecular method detecting variation in length within minisatellite DNA containing variable numbers of tandem repeat sequences. Analysis is by Southern blot hybridization using probes 33.6 and 33.15, M13 phage DNA, or oligonucleotide sequence	Separates individual cell lines across multiple species	Refs. 25,38
Short tandem repeat (STR) profiling	Molecular method detecting variation in length within microsatellite DNA containing variable numbers of short tandem repeat sequences. Analysis is by PCR with comparison to set size standards; usually available in a kit format allowing amplification of up to 16 loci	Separates individual cell lines within a single species	Refs. 26,39
Polymerase chain reaction (PCR) fragment analysis	Molecular method involving amplification of specific genes or gene families, aiming to detect variations in exon/intron sequence, transcript splicing, or the presence of pseudogenes. Genes examined include the aldolase gene family and the beta-globin gene	Separates species only	Refs. 40,41
Sequencing of "DNA barcode" regions	Involves sequencing of a DNA fragment from the mitochondrial gene cytochrome *c* oxidase subunit I, with comparison to sequence obtained from online databases. This "DNA barcode" has been shown in practice to distinguish a broad range of animal species	Separates species only	Refs. 27,42

banks to publish such fingerprints online. It is advisable to always compare the test sample to a known sample within the same experiment, ideally using DNA from the blood or tissue of the original donor.

The obvious advantage of STR profiling lies in the use of control samples to generate a numerical code for each sample, which precisely identifies that cell line and which can be readily shared and published online. It is primarily for this reason that STR profiling is recommended as an international reference standard for human cell lines[26] and accepted within the legal system for human identity testing.[39] STR profiling is based on the presence of STRs within the human genome that exist at variable lengths throughout the population. Each of the repeat regions to be analyzed (usually tetra or pentanucleotide repeats in noncoding sequence) is amplified by PCR using primers carrying fluorescent tags and electrophoresed in a sequencing gel; the precise length of each allele is determined and compared with size standards and controls. This allows identification software to assign a number to each allele at that locus (see, *e.g.*, Fig. 2). The combination of multiple loci—classically 13, as used in the FBI Laboratory's Combined DNA Index System (CODIS)—gives sufficient data to uniquely identify that individual.

STR profiles for individual cell lines and panels have now been reported by many laboratories (*e.g.*, Ref. 44) and are published online by several cell banks. However, there are some cautions to be aware of when using this approach. It is accepted within the forensic field that tumor samples are not as genetically stable as other tissue sources for STR profiling, because of loss of heterozygosity and microsatellite instability.[45,46] This is even more evident in tumor-derived cell lines, where evolution or genetic drift continues to occur with passage.[47] When searching an online database of STR profiles from cell lines, the user needs to look for close matches and not just identical matches; most studies would agree that 80% similarity is an appropriate threshold for declaring a match when comparing cell line profiles.[26,44] There may also be a significant start-up cost if testing in-house; in addition to an STR kit, access to methods for DNA extraction, precise quantitation, fragment analysis and software for STR profile identification is required.

The fact that STR profiling is only suitable for distinguishing cell lines of a single species has led to the need to re-examine authentication of nonhuman cell lines. Laboratory rodent samples will always be difficult to identify precisely due to inbreeding; laboratories working with rat or mouse cultures may wish to examine strain identity rather than authentication of individual cell lines, particularly if they have expertise in single nucleotide polymorphism (SNP) or single sequence length polymorphism (SSLP) analysis,