D512–D515 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh124

Pathbase: a database of mutant mouse pathology

Paul N. Schofield*, Jonathan B. L. Bard[1], Catherine Booth[2], Jacques Boniver[3], Vincenzo Covelli[4], Philippe Delvenne[3], Michele Ellender[5], Wilhelm Engstrom[6], Wolfgang Goessner[7], Michael Gruenberger, Heinz Hoefler[7], John Hopewell[8], Mariatheresa Mancuso[4], Carmel Mothersill[9], Christopher S. Potten[2], Leticia Quintanilla-Fend[7], Bjorn Rozell[10], Hannu Sariola[11], John P. Sundberg[12] and Andrew Ward[13]

Department of Anatomy, University of Cambridge, Downing Street, Cambridge CB2 3DY, UK, [1]Department of Biomedical Sciences, Hugh Robson Building, George Square, Edinburgh University, Edinburgh EH8 9XD, UK, [2]Paterson Institute, The Christie Hospital, Wilmslow Road, Manchester M20 4BX, UK and EpiStem Ltd, Incubator Building, Grafton Street, Manchester M13 9XX, UK, [3]Department of Pathology, University Hospital of Liege, B4000, Liege, Belgium, [4]ENEA, Divisione Protezione dell'uomo e degli Ecosistemi, Via Anguillarese 301, 00060, Rome, Italy, [5]National Radiological Protection Board, Chilton, Didcot, Oxon OX11 0RQ, UK, [6]Department of Pathology, Swedish University of Agricultural Sciences, Uppsala 750 07, Sweden, [7]Institut fuer Pathologie, GSF-Forschungszentrum fuer Umwelt und Gesundheit, Ingolstaedter Landstrasse 1, Neuherberg, 85764, Germany, [8]Department of Clinical Oncology, Churchill Hospital, Oxford OX3 7LJ, UK, [9]Radiation Science Centre, Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland, [10]Clinical Research Center and Division of Pathology, IMPI, Karolinska Institutet, Huddinge University Hospital, SE-141 86 Stockholm, Sweden, [11]Institute of Biomedicine, Developmental Biology, PO Box 63 (Haartmaninkatu 8), FIN-00014 University of Helsinki, Finland, [12]The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA and [13]Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK

Received August 14, 2003; Revised and Accepted October 21, 2003

ABSTRACT

Pathbase is a database that stores images of the abnormal histology associated with spontaneous and induced mutations of both embryonic and adult mice including those produced by transgenesis, targeted mutagenesis and chemical mutagenesis. Images of normal mouse histology and strain-dependent background lesions are also available. The database and the images are publicly accessible (http://www.pathbase.net) and linked by anatomical site, gene and other identifiers to relevant databases; there are also facilities for public comment and record annotation. The database is structured around a novel ontology of mouse disorders (MPATH) and provides high-resolution downloadable images of normal and diseased tissues that are searchable through orthogonal ontologies for pathology, developmental stage, anatomy and gene attributes (GO terms), together with controlled vocabularies for type of genetic manipulation or mutation, genotype and free text annotation for mouse strain and additional attributes. The database is actively curated and data records assessed by pathologists in the Pathbase Consortium before publication. The database interface is designed to have optimal browser and platform compatibility and to interact directly with other web-based mouse genetic resources.

INTRODUCTION

Progress in the functional genomics of the mouse relies heavily on the sophisticated and systematic phenotyping of mutants to help understand the normal function of gene products (1). Such phenotyping depends to a large extent on histopathological analysis of the lesion and pathology is thus a key discipline in this endeavour.

It is, however, very difficult to access the primary phenotypic data on mutant mice. The volume and sheer complexity of the data, together with the numerous links between different components, make it hard to obtain an overview of the relationship between the multiple aspects of a phenotype and the genotype of a mutant. The issue of access to the primary data, usually images, is highly important as sparse descriptive coding alone may miss aspects of the phenotype that subsequently become important. Most current databases of mutant mice use predominantly text rather than images to describe lesions that are part of abnormal phenotypes [e.g. Tbase (2)] while the literature can only offer a few key pictures to supplement text.

Existing image resources focus on specific domains [e.g. RENI (3) and the Jackson Laboratory Tumor Biology

*To whom correspondence should be addressed. Tel: +44 1223 333893; Fax: +44 1223 333754; Email: ps@mole.bio.cam.ac.uk

Database (4,5) cover toxicological pathology and tumour pathology, respectively] and, while these are invaluable, they are often thematically restricted; indeed smaller resources concentrate on the specific domain of interest of, or mutants generated by, the host laboratory. Thus there is a clear need for a community resource that codes, archives and distributes primary experimental data describing mutant mouse lesions. This paper reports the newly available Pathbase database set up by the Pathbase Consortium for this purpose.

In order to provide accurate and 'hypothesis-neutral' ways of retrieving images from the database, Pathbase required the development of a semantic framework for systematically coding images of pathological lesions. As part of this, the Consortium has developed an ontology (MPATH) to describe mouse pathological lesions, described below, which is combined with other orthogonal ontologies and controlled vocabularies (CVs) to describe each image and so enable access and analysis by the user (see Table 1). This ontology has been produced by pathologists and is designed both to act as a terminology reference facility, and to catalogue histopathologic images of mutant mice. The database acts as a repository for such images and users are invited to add images for the benefit of the community. Here, we describe the facilities available in Pathbase and briefly consider the informatics infrastructure of the database.

THE DATABASE INFRASTRUCTURE

Ontologies and meta-data

A key challenge in the construction of Pathbase has been the development of approaches to the coding of images. The semantic meta-data for each Pathbase image consists of the relevant terms from a set of orthogonal ontologies and short CVs that have recently been developed by us and other groups (6). The anatomical attributes of the image are coded by using either the time-dependent mouse embryo anatomy developed by the EMAGE project (7,8) or the adult anatomy developed by the Gene Expression Database (GXD) project (9) at the Jackson Laboratory. Other attributes are coded using the pathology (MPATH) ontology and a series of short CVs. Allele names and strain nomenclature are in accordance with the MGD gene annotation group following the International Committee on Standardized Genetic Nomenclature for mice (10) and records are associated directly with the Gene Ontology, or GO, terms (11) for the mutant gene or genes, assigned by MGI (12). All these ontologies, which are publicly available from source files at http://obo.sourceforge.net, are used for both archiving data and querying the database.

MPATH itself is a new, actively curated, fully defined ontology that contains terms (currently 555) covering all currently known classes of lesion, with specific reference to the mouse, arranged as a hierarchy. The inclusion of definitions and synonyms helps to clarify the often disparate set of terms used by pathologists trained in different traditions which actually describe the same lesion. It incorporates the NIH Mouse Models of Human Cancer Consortium recommendations on haematopoietic neoplasms (13,14) and will be maintained in accordance with existing reviews of other mouse cancer types from this body, and new ones as they are published.

The ontology is accessible both from Pathbase and from the GO Consortium's Open Biological Ontology (OBO) site (11,15) where bio-ontologies are archived (http://obo.sourceforge.net). Technically, the MPATH hierarchy is held within a directed acyclic graph (DAG) six levels deep, written in DAGEDIT format and instantiated in GO syntax. Each item in the ontology has an MPATH ID that can be used for analysis and database interoperability.

The meta-data for each record incorporates the appropriate OBO IDs and this aspect allows Pathbase to be easily interoperable with other databases. It is also worth noting that an image may have multiple terms assigned to it from the same ontology, where, for example, there are elements of neoplasia and metaplastic change in the same image. All ontologies used on Pathbase are updated regularly and automatically from source files at http://obo.sourceforge.net and a system has been implemented using standard software to facilitate this.

Implementation

The summary of how Pathbase works is taken from technical notes available on the website (http://www.pathbase.net/help/developers). In brief, the system is based on four layers, with the user only seeing that used for presentation. Underlying this are database, application and transport layers.

Database layer. The system is currently using Sybase ASE 11.9.2 as a database back-end within a Gentoo Linux environment. The Pathbase database schema is divided into two sections: one handling images and the other the meta-data structured through the ontologies.

Application layer (PHP). This layer formulates SQL queries from HTTP input and formats search results as HTML, XML, etc. The infrastructure set-up will allow for seamless transition to another database engine in future (e.g. MySQL or PostgreSQL).

Transport layer. A standard installation of Apache 1.3.27 handles this layer, using routine compression software suitable for users with slow internet connection and for overseas users. The web server also handles some of the security of the website, e.g. the password protection of the administration area.

Presentation layer. This layer is the user's web browser or another database server. Pathbase sends standard HTML and CSS, and uses standard JavaScript for the pop-up ontology windows. We are currently implementing a DHTML help system that displays instructions when users mouse-over input fields.

The ontologies themselves are stored as flat files and converted to the required formats and hierarchies on-the-fly through PHP scripts. The same ontologies and CVs are used combinatorially for searching and for archiving data.

Searching the database can most simply be done through the ontologies and CVs which are incorporated into the user interface. For the ontologies, searches are carried out hierarchically using all of the terms below that selected. However, as there is free text associated with the annotations of the images, there is also a string search facility that allows a user to search the database through user-generated keywords.

Table 1. Fields describing images in the database

Description (mandatory field)	Items	Description (non-mandatory field)	Items
Pathology (ontology)	MPATH	Embryonic age (ontology)	EMAP
Post-natal anatomy (ontology)	MA	Post-natal stage (CV)	6
Genetic manipulation (CV)	21	Gene	Free text
Strain	Free text	Allele name	Free text
Genotype status (CV)	7	Gene Ontology	GO function
			GO process
			GO component
Organism (CV)	3	TBase identifier	Free text
Sex (CV)	3	Medline identifier	Free text
Description	Free text	Stain	Free text
		Magnification	Free text
		Experimental manipulation (in addition to original mutation) (CV)	5

Image submission requires information for the mandatory fields while the others are non-essential or may be inappropriate for some datasets. Fields utilize short controlled vocabularies (CVs), free text or a detailed ontology. All fields can be used for searching.

REGISTRATION AND ACCESS

Registration is only required for submissions (to ensure a high quality of data and for copyright reasons) and can be done through the database administration or directly by the user on the web. Registration and use are free to all users, and are not required for searching the database and retrieving images.

DATA RETRIEVAL

The core of Pathbase is a set of images and associated meta-data that can be searched via the ontologies, CVs or free text. More than 1000 images are currently on line at the time of submission. These images are mainly from mutant mice (e.g. transgenics, targeted mutants, chemical- and radiation-induced and spontaneous mutations); however, images of normal histology and strain-dependent background lesions are also available.

Pathbase images and associated meta-data can be remotely searched via the ontologies, CVs and free text. A request yields thumbnails of all appropriate images together with appropriate data and these thumbnails may be expanded on the screen to give full-screen JPG images. These JPG files may be downloaded to the user's computer but any associated TIFF files held within the database will only be available on request due to the large uncompressed file size.

DATA ACQUISITION AND ANNOTATION

Data acquisition proceeds in two modes. The curatorial team actively request published images or additional unpublished material from existing papers, from existing image collections or from groups generating large amounts of phenotype data, such as ENU mutagenesis centres. Alternatively, users may send or upload their own images via user-friendly interfaces, FTP transfer, email or portable media such as CDs. Images can be uploaded in all common formats with an optimal resolution of at least 300 pixels per inch and a maximum file size of 8 Mbyte. The Pathbase Consortium is also able to accept transparencies and slides, which will be scanned and returned.

Sending images to the database requires that the submitter also send key associated data and the entry interface page for this has been made as simple as possible through the use of the ontologies and CVs, which merely require the user to click on an appropriate term. Searching for images uses the same ontologies and CVs and is again designed to be simple.

Data records are linked to other core web resources. In many cases these records will carry a Tbase identifier (2), linking the two resources. Image records may also be linked to MGD, the mouse genome database (16), while literature may be accessed through PubMed. We see the interoperability of Pathbase with other related databases as an area we wish to develop strongly in the future.

Following upload, all records are subject to active curation and assessment by a member of the Pathbase Pathology Committee before going online. A curatorial interface has been developed which is related to the publicly available upload interface but has facilities for adding curatorial notes and data reliability level; these provide important traceability of the annotation. The geographical dissemination of the pathologists in the Consortium makes such an interface extremely valuable during assessment of the records.

Pathbase also allows users to annotate image records directly providing a route for direct user feedback. Although moderated by the curators, such additions to the information on the records will build up a community expertise resource not so far attempted in a database of this kind.

DISCUSSION

The first phase in the evolution of biological databases covered the archiving of sequence data in its many forms. The next stage was the production of databases that unified particular groups by providing them with a community resource [e.g. Flybase (17)]. An important current need is for databases that carry image data that provides visual as well as textual data and that is linked both to formal knowledge systems (ontologies) and to other core databases; Pathbase is one of a group of such databases that is now available. In terms of informatics, one notable feature of Pathbase is its use of formal ontologies for searching, and considerable effort has gone into making its pathology ontology (MPATH) as comprehensive and authoritative as possible.

The Pathbase ontology is not intended as an alternative diagnostic framework to the existing nomenclatures such as SNOVET (18), but as a tool for data retrieval that provides a

set of inclusive terms into which all pathological lesions can be fitted. SNOVET now contains over 100 000 terms and is not appropriately structured to describe the pathobiology of mutant mice as many of its terms are dependent on aetiology and anatomical location. Whilst the 'lumping' of some terms in the MPATH ontology does remove some diagnostic precision, this has the advantage that small differences in opinion as to the precise diagnosis, or usage of terms, which vary between different traditions of pathology, do not affect the accuracy of a search. Such precise diagnostic terms can still be entered into the free text field and searched independently.

It is important to emphasize that MPATH is not in itself a phenotype or a disease ontology, it is designed as a description ontology to provide the key meta-data of images of lesions in tissues generated in response to underlying genetic or extrinsic damage. There is currently much debate concerning the best way of coding complete phenotype data: either by a series of orthogonal ontologies or by description-logic-based systems using ontologies as plug-ins (19). Both approaches are still at an early stage and development of an ontology for comprehensively describing 'disease' phenotypes for the mouse has not yet been attempted, the closest being that recently developed by the MGD project (16) at the Jackson Laboratory (MPheno; available on http://obo.sourceforge.net), which uses a combination of defined anatomical, pathobiological and cellular terms from within a single ontology to summarize the phenotype.

Pathbase itself is still only a facility. While it holds many images and their associated information, data acquisition is an open-ended process (unlike an ontology which is intended to encapsulate a well-defined domain of knowledge). Its success will depend on users being prepared to use the facility to store their own data and thus on a certain degree of altruism. To help here, the curators have tried to make it as easy as possible for users to upload or send images and hope that users who are publishing interesting pictures of mutation-associated abnormalities in both embryonic and adult rodents will also submit them to the database. Pictures that are submitted by users will be rapidly made available to the field. New images and additions to the literature found by the curators will be entered into the database and, if appropriate, added to a future reference site.

Pathbase is intended to have two uses. The first is to help biologists with a limited knowledge of mouse pathology to assess and compare their data, and to provide access to more expert advice. The second is to allow the mouse community to share its knowledge of image-based pathology. The first of these aims should be met by the curation team, the latter requires an involvement by the user community. We invite you, the reader, to take advantage of this opportunity.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Jerry Ward and Dr Janan Eppig for their support during the development of Pathbase. The project is funded by the European Commission under Framework 5, contract number QLRI-1999-CT-00320.

REFERENCES

1. Nolan,P.M., Peters,J., Strivens,M., Rogers,D., Hagan,J., Spurr,N., Gray,I.C., Vizor,L., Brooker,D., Whitehill,E. *et al.* (2000) A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nature Genet.*, **25**, 440–443.
2. Anagnostopoulos,A.V., Mobraaten,L.E., Sharp,J.J. and Davisson,M.T. (2001) Transgenic and knockout databases: behavioral profiles of mouse mutants. *Physiol. Behav.*, **73**, 675–689.
3. Mohr,U. (ed.) (2001) *International Classification of Rodent Tumors. The Mouse*. Springer Verlag, Berlin (http://www.item.fraunhofer.de/reni/index.htm)
4. Bult,C.J., Krupke,D.M., Naf,D., Sundberg,J.P. and Eppig,J.T. (2001) Web-based access to mouse models of human cancers: the Mouse Tumor Biology (MTB) Database. *Nucleic Acids Res.*, **29**, 95–97.
5. Naf,D., Krupke,D.M., Sundberg,J.P., Eppig,J.T. and Bult,C.J. (2002) The Mouse Tumor Biology Database: a public resource for cancer genetics and pathology of the mouse. *Cancer Res.*, **62**, 1235–1240.
6. Bard,J. (2003) Ontologies: Formalising biological knowledge for bioinformatics *Bioessays*, **25**, 501–506.
7. Bard,J.L., Kaufman,M.H., Dubreuil,C., Brune,R.M., Burger,A., Baldock,R.A. and Davidson,D.R. (1998) An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. *Mech. Dev.*, **74**, 111–120.
8. Bard,J. and Winter,R. (2001) Ontologies of developmental anatomy: their current and future roles. *Brief. Bioinform.*, **2**, 289–299.
9. Ringwald,M., Eppig,J.T., Begley,D.A., Corradi,J.P., McCright,I.J., Hayamizu,T.F., Hill,D.P., Kadin,J.A. and Richardson,J.E. (2001) The Mouse Gene Expression Database (GXD). *Nucleic Acids Res.*, **29**, 98–101.
10. Maltais,L.J., Blake,J.A., Chu,T., Lutz,C.M., Eppig,J.T. and Jackson,I. (2002) Rules and guidelines for mouse gene, allele and mutation nomenclature: a condensed version. *Genomics*, **79**, 471–474.
11. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* the Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
12. Hill,D.P., Davis,A.P., Richardson,J.E., Corradi,J.P., Ringwald,M., Eppig,J.T. and Blake,J.A. (2001) Program description: Strategies for biological annotation of mammalian systems: implementing gene ontologies in mouse genome informatics. *Genomics*, **74**, 121–128.
13. Kogan,S.C., Ward,J.M., Anver,M.R., Berman,J.J., Brayton,C., Cardiff,R.D., Carter,J.S., de Coronado,S., Downing,J.R., Fredrickson,T.N. *et al.* (2002) Bethesda proposals for classification of nonlymphoid hematopoietic neoplasms in mice. *Blood*, **100**, 238–245.
14. Morse,H.C.,3rd, Anver,M.R., Fredrickson,T.N., Haines,D.C., Harris,A.W., Harris,N.L., Jaffe,E.S., Kogan,S.C., MacLennan,I.C., Pattengale,P.K. *et al.* (2002) Bethesda proposals for classification of lymphoid neoplasms in mice. *Blood*, **100**, 246–258.
15. GO Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
16. Blake,J.A., Richardson,J.E., Bult,C.J., Kadin,J.A. and Eppig,J.T. (2003) MGD: the Mouse Genome Database *Nucleic Acids Res.*, **31**, 193–195.
17. Flybase Consortium (2003) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.*, **31**, 172–175.
18. Palotay,J.L. (1983) SNOMED-SNOVET: an information system for comparative medicine. *Med. Inform. (Lond.)*, **8**, 17–21.
19. Hill,D.P., Blake,J.A., Richardson,J.E. and Ringwald,M. (2002) Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. *Genome Res.*, **12**, 1982–1991.