PATHBASE: A NEW REFERENCE RESOURCE AND DATABASE FOR LABORATORY MOUSE PATHOLOGY

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

[1]Department of Anatomy, University of Cambridge, Downing Street, Cambridge CB2 3DY, UK
[2]Department of Biomedical Sciences, Hugh Robson Building, George Square, Edinburgh University, Edinburgh EH8 9XD, 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]Applied Radiation Sciences Unit, McMaster University, Nuclear Research Building, Rm 228, 1280 Main Street West, Hamilton, L8S 4K1, Ontario, Canada
[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
[13]Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK

Pathbase (http://www.pathbase.net) is a web accessible database of histopathological images of laboratory mice, developed as a resource for the coding and archiving of data derived from the analysis of mutant or genetically engineered mice and their background strains. The metadata for the images, which allows retrieval and interoperability with other databases, is derived from a series of orthogonal ontologies and controlled vocabularies. One of these controlled vocabularies, MPATH, was developed by the Pathbase Consortium as a formal description of the content of mouse histopathological images. The database currently has over 1000 images on-line with 2000 more under curation and presents a paradigm for the development of future databases dedicated to aspects of experimental biology.

INTRODUCTION

With the completion of its genomic sequence, the mouse has become *par excellence* the prime surrogate organism for the investigation of human biology and disease[1]. The success of the 'post-genomic' phase of the mouse genome project now depends critically on the analysis of the phenotypes associated with spontaneous and experimentally induced genetic lesions. Through detailed phenotype analysis, including the response to challenges such as ionising radiation, it is possible to establish gene function in the intact organism, develop mouse models of human disease and gain an understanding of common normal developmental and physiological processes (e.g. see Mancuso *et al.*[2]). The volume of data, together with the detailed analysis of existing mutants will soon, if it is not already, be overwhelming and much high quality primary data risks being lost as only selected highlights end up in the literature, often in an inaccessible and unstructured way. Moreover, this data is distributed across databases that are not interoperable so that it is difficult for a user to integrate the set of answers that each database will give to a specific query.

Systematic pathological analysis is a key requirement for the understanding of mutant phenotypes and the response of such animals to experimental manipulation. It was primarily to meet this need that Pathbase (http://www.pathbase.net) was created[3] with support from the European Commission in order to support the histopathological analysis and dissemination of data derived from mutant mice. A serious problem, currently faced by the community, is the worldwide

*Corresponding author: PS@mole.bio.cam.ac.uk

lack of available pathology expertise to carry out these analyses, with relatively few pathologists having experience of the distinctive pathology and genetics of mutant laboratory mice. Consequently, one of the secondary aims of Pathbase was to provide a reference resource to aid education and diagnosis.

In this article, the development and use of Pathbase is described and in addition, the problems associated with the description of mouse lesions and database interoperability is examined.

Pathbase informatics

Pathbase set out to provide a means for archiving and retrieving the large amounts of pathology data that are associated with, in particular, mutant and laboratory strains of mice, the first task being the development of a framework or systematised nomenclature for coding images of all the major pathological lesions. For this, a group of experienced pathologists and geneticists, the Pathbase Consortium, have collated all the major classes of mouse pathology and organised them within a formal subsumption hierarchy (or ontology). These terms, together with an anatomical ontology and a set of controlled vocabularies (CVs), can be used for classifying, archiving and retrieving the original images. The database has also been established using a format that lends itself to interoperability with other databases.

Pathbase (http://www.pathbase.net) now has >3000 images that are currently being coded with associated metadata linked to ontologies and CVs

(see below). The first public release with over 1000 images was in June 2003 and the next major release will take place late 2004 with another 2000 images.

ONTOLOGIES AND METADATA

A key aspect of Pathbase has been the development and implementation of formal descriptive metadata for pathological images. Pathbase is neither gene-, nor overall phenotype-centric and therefore does not aim to hold complete systematic descriptions of the phenotype associated with individual mutations or manipulations; it does, nevertheless, allow a user to retrieve images on the basis of, e.g. the underlying mutation, strain, anatomical location or pathological lesion. The metadata for each image within Pathbase consists of the relevant terms from a set of orthogonal ontologies and short CVs, recently developed by this and other groups[4]. These are shown diagrammatically in Figure 1. This metadata makes Pathbase a powerful tool for discovery and increases its value as a specialised repository from which other gene- or phenotype-centric databases may obtain information. Database interoperability was a key factor in the development of the Pathbase data structure and, with the setting up of web services for other databases, the value of this is now being fully realised.

The anatomical attributes of the image are coded by using either the time-dependent embryo anatomy ontology developed by the EMAGE consortium (http://genex.hgu.mrc.ac.uk)[5,6] or the mouse adult
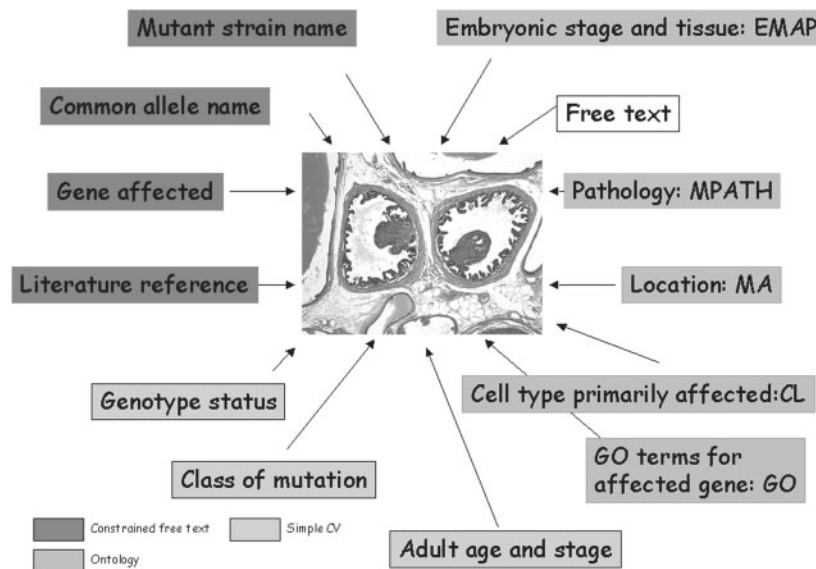

Figure 1. Diagrammatic representation of the metadata classes associated with each image in Pathbase.

anatomy (MA: http://www.informatics.jax.org/searches/anatdict_form.shtml) developed for the gene-expression database (GXD) initiated by and curated at the Jackson Laboratory (http://www.informatics.jax.org/mgihome/GXD/about GXD.shtml). Other attributes are coded using the cell type (CA: Ashburner and Bard. Pers. Comm.) and pathology (MPATH) ontologies. Allele names and strain nomenclature are in accordance with that curated by MGD[7]. In addition, the records are directly associated with the Gene Ontology (GO) terms[8] of the mutant gene, assigned by MGD[7] to provide data on the cell location and molecular function of the gene as well as the biological process in which the gene is involved. Therefore, each image is defined by the intersection of a series of CV, ontology and free text terms that are used for both archiving data and querying the database. In addition, the concept of a 'constrained free text' term, where it is not feasible to maintain a defined ontology or CV, has been introduced. These are used for gene and allele names, which are continuously under review, and for those names that are defined by an authority (e.g. the Mouse Genome Informatics group who provide a curated reference resource).

MPATH is a new, curated and fully defined ontology that contains terms covering all currently known classes (580) and instances of pathology, with specific reference to the mouse, arranged as a hierarchy within a Directed Acyclic Graph (DAG), six levels deep. Each item has an MPATH ID that can be used for database interoperability and analysis; the ontology is accessible both from the Pathbase website and from the Gene Ontology Consortium's OBO (Open Biological Ontologies) site[8,9] where bio-ontologies are archived (http://obo.sourceforge.net). MPATH has recently been restructured to take account of the MMHCC (Mouse Models of Human Cancer Consortium) recommendations on haematopoietic neoplasms[10,11] and will be maintained in accordance with reviews on other mouse cancers from this body as they are published (http://emice.nci.nih.gov/emice/mouse_models).

It is important to emphasise that MPATH is not a phenotype or a disease ontology, it is a descriptive ontology for handling the key metadata of images of lesions in tissues generated in response to underlying genetic or extrinsic damage, however, images may be described by a combination of terms referring to content which include both general and specialised pathology terminology.

Using Pathbase

The core of Pathbase is a set of images and associated metadata that can be searched via the ontologies (or free text). These images are mainly from mutant mice (e.g. transgenics, knockouts), chemical- and radiation-induced and spontaneous mutations. However, a new resource from the Jackson Laboratory will define spontaneous, strain specific diseases in commonly used strains and stocks of laboratory mice. Most of the images in Pathbase to date, have come from major international databases following requests from the consortium and from individuals. Researchers may also directly deposit their own images (with associated metadata) via user-friendly interfaces, ftp transfer, email or portable media such as CD/DVDs. 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 25 MB.

Searching of the database can be carried out from an interface which provides open fields for free text, drop down boxes for short CVs, and pop-up boxes in which the ontologies may be searched and terms transferred to the search field with a single mouse click. Combinations of terms may be used in queries and, e.g. it is possible to search for lesions resulting from the mutation of functional classes of gene (rather than specific individual genes) as specified by the GO terms associated with them, in tissues of interest. Images in the database may be downloaded as JPG files, with original TIFFs available on request. Copyright of images remains with the originators and an assertion to that effect is explicit on the web site, with adherence a condition of use. TIFFs will only be provided to users with agreement of the original donor.

EM*pathy*

This database provides a reference and teaching resource for mouse pathology. The key feature of the reference component is the MPATH ontology of known pathological lesions that is linked to a set of reference images (>50% complete). EM*pathy* also holds other useful resources such as an illustrated and downloadable guide to conducting a mouse necropsy. The web site is organised by organ and tissue, and shows users the abnormalities associated with each anatomical location. Each pathology record consists of a set of reference images and notes on the aetiology, diagnostic criteria and, where appropriate, reference to the background incidence of the disorder and examples of the occurrence in genetically modified mice. There are currently about 200 such lesions listed and we intend to double the number over the next year. EM*pathy* is due to go online before the end of 2004.

A VISION FOR FUTURE INTEGRATION OF RADIOBIOLOGY DATA

The establishment of radiobiology databases has happened much more slowly than those designed to serve functional genomics. Pathbase has a considerable amount of pathology data resulting from

radiation exposure and ingestion in rodents and coding of these images means that experimental radiation databases, such as ERA (European Radio-biology Archive; http://www.gsf.de/ERA), would in principle be able to link to Pathbase for useful information, if its pathology codes could be cross matched to the MPATH codes. Similarly adoption of the mouse adult anatomy ontology, MA and phenotype ontologies as they are developed and deposited on OBO would enhance interoperability still further.

Pathbase is a model for future database interaction and exemplifies a paradigm for the future of radiobiology databases. A great deal of model organism data currently exists in repositories in Europe and the US, which needs to be cross referenced to already existing human data and that being accumulated through ongoing studies. The adoption of the same description frameworks for this data, specifically in terms of radiobiology and pathology would provide enormous added value for the comparison of mouse and human models of radiation response; the challenge to the community is to agree on standards and generate common tools that will augment our effectiveness in research.

ACKNOWLEDGEMENTS

REFERENCES

1. Sands, A. T. *The master mammal*. Nat. Biotechnol. **21**, 31–32 (2003).
2. Mancuso, M. *et al. Basal cell carcinoma and its development: insights from radiation-induced tumors in Ptch1-deficient mice*. Cancer Res. **64**, 934–941 (2004).
3. Schofield, P. N. *et al. Pathbase: a database of mutant mouse pathology*. Nucleic Acids Res. **32** Database issue, D512–515 (2004).
4. Bard, J. B. and Rhee, S. Y. *Ontologies in biology: design, applications and future challenges*. Nat. Rev. Genet. **5**, 213–222 (2004).
5. Bard, J. L. *et al. An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature*. Mech. Dev. **74**, 111–120 (1998).
6. Bard, J. and Winter, R. *Ontologies of developmental anatomy: their current and future roles*. Brief Bioinform. **2**, 289–299 (2001).
7. Hill, D. P. *et al. Program description: strategies for biological annotation of mammalian systems: implementing gene ontologies in mouse genome informatics*. Genomics **74**, 121–128 (2001).
8. Ashburner, M. *et al. Gene ontology: tool for the unification of biology*. The Gene Ontology Consortium. Nat. Genet. **25**, 25–29 (2000).
9. The Gene Ontology Consortium, *Creating the gene ontology resource: design and implementation*. Genome Res. **11**, 1425–1433 (2001).
10. Morse, H. C., 3rd *et al. B lymphoid neoplasms of mice: characteristics of naturally occurring and engineered diseases and relationships to human disorders*. Adv. Immunol. **81**, 97–121 (2003).
11. Kogan, S. C. *et al. Bethesda proposals for classification of nonlymphoid hematopoietic neoplasms in mice*. Blood **100**, 238–245 (2002).