Published online 23 October 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D593–D599***
doi:10.1093/nar/gkp867

MouseBook: an integrated portal of mouse resources

Andrew Blake, Karen Pickford, Simon Greenaway, Steve Thomas, Amanda Pickard, Christine M. Williamson, Niels C. Adams, Alison Walling, Tim Beck, Martin Fray, Jo Peters, Tom Weaver, Steve D. M. Brown, John M. Hancock and Ann-Marie Mallon*

MRC Harwell, Mammalian Genetics Unit and the Mary Lyon Centre, Harwell Science and Innovation Campus, Oxfordshire, OX11 0RD, UK

Received August 14, 2009; Accepted September 28, 2009

ABSTRACT

The MouseBook (http://www.mousebook.org) databases and web portal provide access to information about mutant mouse lines held as live or cryopreserved stocks at MRC Harwell. The MouseBook portal integrates curated information from the MRC Harwell stock resource, and other Harwell databases, with information from external data resources to provide value-added information above and beyond what is available through other routes such as International Mouse Stain Resource (IMSR). MouseBook can be searched either using an intuitive Google style free text search or using the Mammalian Phenotype (MP) ontology tree structure. Text searches can be on gene, allele, strain identifier (e.g. MGI ID) or phenotype term and are assisted by automatic recognition of term types and autocompletion of gene and allele names covered by the database. Results are returned in a tabbed format providing categorized results identified from each of the catalogs in MouseBook. Individual result lines from each catalog include information on gene, allele, chromosomal location and phenotype, and provide a simple click-through link to further information as well as ordering the strain. The infrastructure underlying MouseBook has been designed to be extensible, allowing additional data sources to be added and enabling other sites to make their data directly available through MouseBook.

INTRODUCTION

The mouse plays a fundamental role in the study of mammalian biology and human disease (1). Since completion of the sequencing of the mouse genome in 2002 (2), there has been great emphasis on its use in conjunction with mutant mice and their identifiable phenotypes to understand the functional significance of the individual genes in the mouse and human genomes and their relationship to disease (3). This has led to a massive proliferation of different kinds of databases dealing with mouse genotype and phenotype information (4) and consequent difficulties for individual lab-based researchers in identifying resources relevant to their research.

The general proliferation of databases has necessitated the development of new approaches for integrating and accessing the data they contain. A popular idea in the biosciences is the idea of a 'bioinformatics nation' (5–7), wherein many databases are linked by providing computational access via web services. This allows the mining of data from multiple databases by a single portal (8). Portals based around core databases but bringing in data from other related databases in real time via web services may be the way forward in providing easy access to diverse datasets.

MouseBook (http://www.mousebook.org) seeks to take advantage of this approach in a particular context. MRC Harwell is a major provider of cryopreserved mutant and inbred mouse strains via its Frozen Embryo and Sperm Archive (FESA) core as well as being a holder of a number of unique scientific data sources such as the Imprinting Catalog. Information about the mouse strains is available through the International Mouse Stain Resource (IMSR) website (9), but the information presented through IMSR is relatively sparse. MouseBook has therefore been designed to integrate information held in MRC Harwell's in-house databases, which underlie the core functionality of MouseBook and are manually curated to ensure accurate nomenclature, with information held at other sources regarding relevant genotype and phenotype information. The aim of this value-added approach is to make search for mouse lines of interest easier and to provide a richer information source so that individual lines can be evaluated for their potential

*To whom correspondence should be addressed. Tel: +44 01235 525278; Fax: +44 01235 841210; Email: a.mallon@har.mrc.ac.uk

usefulness. This will have the benefit of increasing the efficiency of mouse research in the face of proliferating numbers of new mouse lines, especially as the International Knockout Mouse Project (IKMC) (10) bears fruit, and will have benefits for the replacement, refinement and reduction of animal research (11).

In the longer term, MouseBook is being developed as a portal through which other laboratories, which may not have MRC Harwell's informatics infrastructure, may present information on mouse lines they wish to publicize making use of MRC Harwell's information management systems, and as a portal to relevant services that can be provided at Harwell and other sites.

DATA IN MouseBook

The MouseBook portal draws data from a number of independent data sources using the architecture described in the implementation section (Figure 1), and organizes these data into bins termed as 'catalogs'. A catalog is a collection of data about a specific entity such as mutant mouse strains (e.g. the Mouse Catalog), with common data elements such as genes integrating the catalogs together to ensure that the data are presented in a user-friendly manner to the portal. The major catalogs within MouseBook are Mouse Catalog, Imprinting Catalog and Chromosomal AnomaliesCatalog.

The Mouse Catalog is an integrated catalog of mutant, and inbred-mouse strains are available from a number of resources. Currently, the mouse strain resources that the Mouse Catalog can access are the MRC Harwell Resource and the European Mouse Mutant Archive (EMMA) Resource (12); however in the future, this catalog could include information from international efforts such as the IKMC or from smaller research laboratories. The MRC Harwell resource is a collection of mutant strains that have been archived at Harwell by the FESA core from the mid-1970s to protect valuable mouse strains against breeding failure, catastrophic losses, genetic drift and genetic contamination while eliminating the need to maintain breeding colonies that are not part of an active research program. In addition, spermatozoa have been archived from more than 10 000 F1 males generated within Harwell's ENU mutagenesis program and a DNA archive established.

At Harwell, these data are captured into the StockList Information Management System (Figure 2), which is an open source standalone application allowing curators to directly enter, query and modify the information associated with an entry. This application also exports newly added and curated information to the IMSR database. Mutant strains can only be identified correctly by users of MouseBook if the information describing them is as accurate as possible. The required information that is crucial to a user identifying the mouse and is of highest priority to be manually curated is the official strain name, the synonym or common name, the genetic background of the mutation, the type of mutation, the affected gene(s) and/or alleles and the phenotypic descriptions.

The StockList Information Management System downloads data into Mouse Genome Informatics (MGI) genes and alleles from the Mouse Genome Database (MGD) (13) FTP reports, and integrates these within the interface to enable the curators to pick from a pull-down

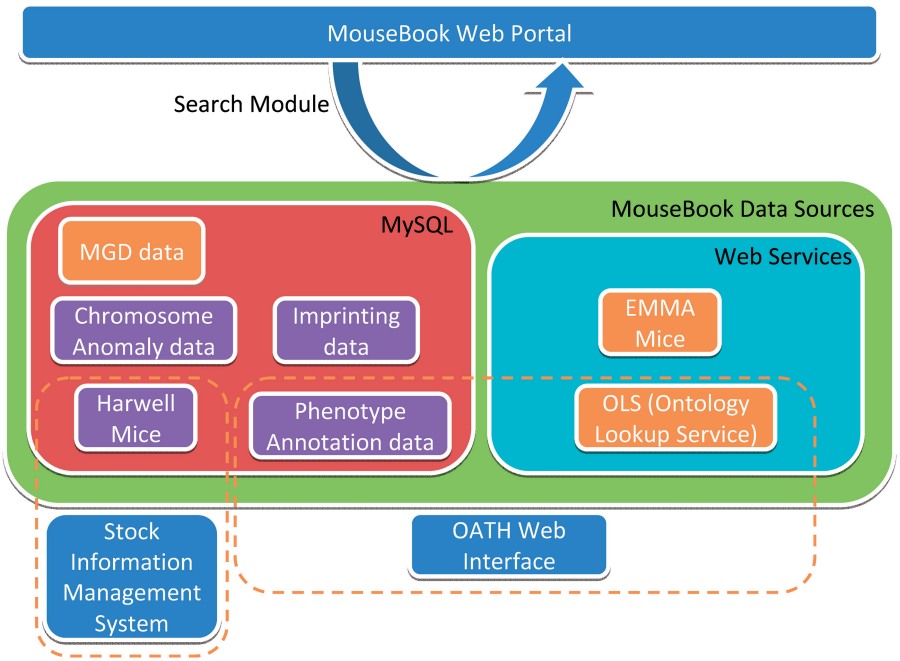

Figure 1. Architecture overview showing how the MouseBook web portal interacts with MouseBook Data Sources. Internal sources (purple boxes) are interrogated using SQL queries. External data sources (orange boxes) are either interrogated using web services or, in the case of MGI data, imported into an internal MySQL database from reports on the MGI FTP site. OATH and the Stock Information Management System interact with the relevant databases to add functionality.

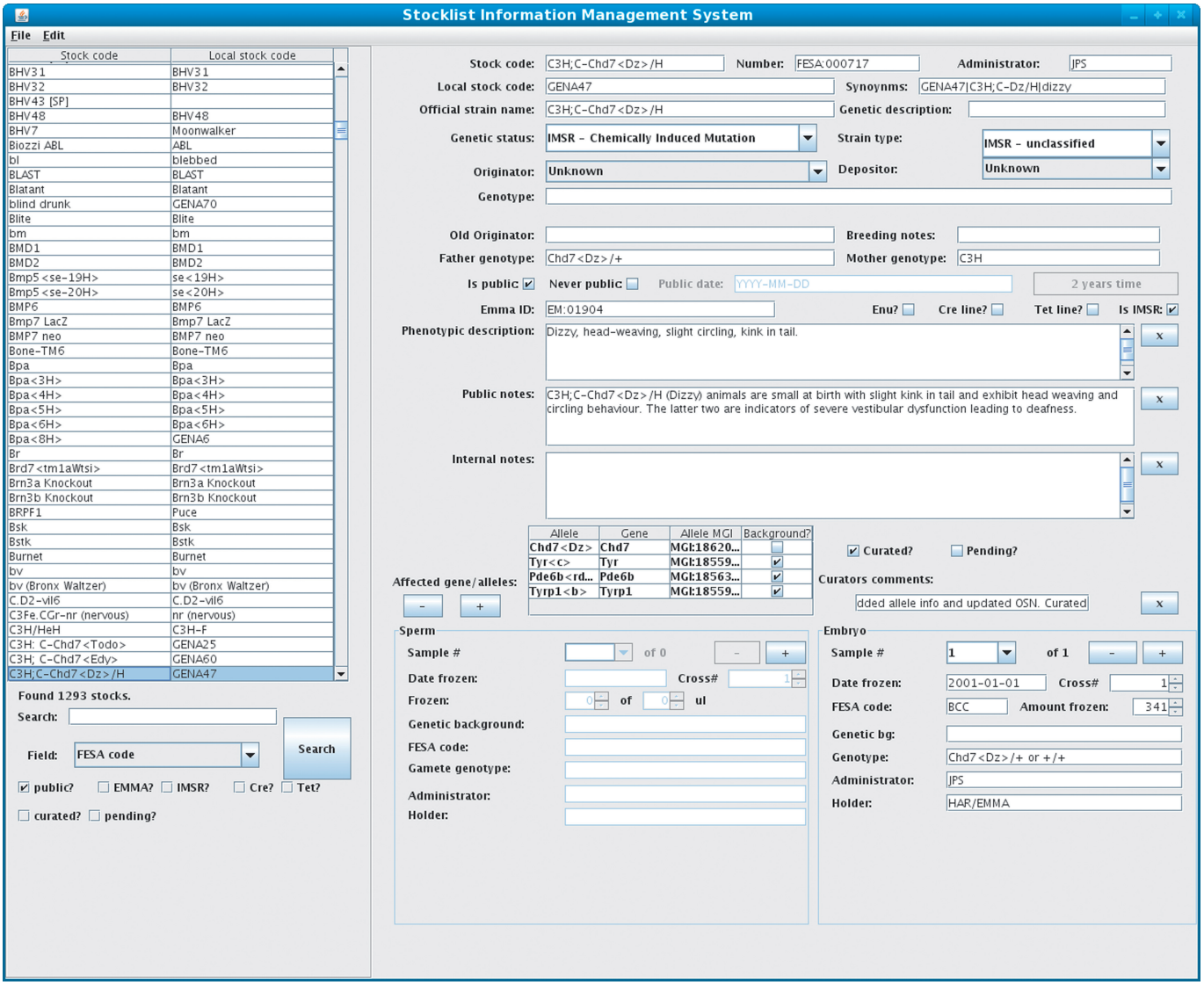

Figure 2. StockList Information Management System interface. This interface enables curators of the MRC Harwell Resource to directly enter new mutant strains or query and modify existing information. The screenshot shows the list of all strains on the left-hand side. Clicking on one stock, e.g. C3H;C-Chd7<Dz>/H, shows all the current information about it. Each of the boxes can then be edited directly. Adding a new affected gene/allele launches a pop-up box enabling the user to select the MGI gene/allele they require. The interface also captures information such as whether the stock is in IMSR or EMMA. The bottom of the interface enables curators to store information which links to the Harwell internal animal husbandry LIMS system, Anonymus, which stores data about the physical samples.

list of genes and alleles to ensure that the mutant strains in the database are assigned an MGI gene and/or allele ID as well as the name. For alleles that are not registered with MGI, the tool allows the curator to add a proposed allele name which will subsequently be registered with MGI. This feature ensures that through MouseBook, the Harwell data can be integrated with more exhaustive information about the affected gene and/or allele from MGI.

The phenotypic descriptions currently captured with a mutant strain are free-text descriptions submitted by the originator. Free text descriptions are difficult to search from a web portal as the search will be unspecific. For example, a search for 'increased body weight' would not identify mutants annotated as 'obese' or 'overweight'. Additionally, subsequent phenotypic information may be identified and published about the mutant, which may not have been known at the time of submission. To ensure users can identify these mutants, they are annotated using the appropriate phenotype ontologies such as the Mammalian Phenotype (MP) Ontology (14). As the amount of high-throughput phenotyping data of mouse mutants from projects such as EUMODIC (http://www.eumodic.org) is increasing in databases such as EuroPhenome (15), annotation of existing mutants with MP terms also ensures that data integration across all mouse mutants can occur at the phenotypic level.

The curation of free-text descriptions is a long and time-consuming task, so the free-text descriptions held in the Stocklist Information Management System are exported into a tool called Ontology Annotation sysTem at Harwell (OATH) (A. Blake *et al.*, manuscript in preparation),

which automatically scans the free text for MP or PATO (16) ontology terms, giving the curator the most likely terms for that text. The mutant strain can then be given the appropriate terms and their ID. Currently this curation is at an early stage.

The EMMA resource is a collection of mutant strains from the European Mouse Mutant Archive, which is an international infrastructure for archiving and distributing mouse mutant strains. MouseBook accesses the data in the EMMA database as described in the implementation section.

Finally, MouseBook provides access to some specialized resources hosted at Harwell and overseen by Harwell scientists. Primary amongst these are the Harwell Imprinting Catalog and Chromosome Anomalies Catalog. The Imprinting Catalog contains information on mouse chromosomal regions associated with imprinted phenotypes, imprinted genes within these regions and imprinted genes in other regions of the genome. The Chromosome Anomalies Catalog contains information on aneuploids and structural rearrangements and their effects.

PORTAL FUNCTIONALITY

In order for the user to be able to search through all of MouseBook, several different query methods have been developed. The aim of the MouseBook search is to make it very simple to use yet flexible, and powerful enough to search MouseBook efficiently, providing the most relevant results to the user quickly.

MouseBook's main search interface is a simple Google style free text search, which enables the user to easily start searching through the data held within MouseBook. The free text search string is text matched against all data held within the MouseBook Catalogs building up a list of results (Figure 3). To make the search more powerful, the free text is also scanned to recognize common identifiers (e.g. 'MGI:' or 'EMMA:') as well as known gene/allele symbols and their synonyms that are then automatically used alongside the text search to provide the user with more accurate and relevant search results. Information from MGD is automatically integrated at the database level within MouseBook, allowing the user to leverage additional information in their search such as marker reference sequence identifiers, cM position, synonyms, protein identifiers, etc.

The results present the primary data with integrated links to information from MGD, Ensembl (17), OMIM (Online Mendelian Inheritance in Man) (18), EMMA and IMSR, where relevant and in the case of information from the Mouse Catalogs it provides the facility to order the cryogenic material.

Alongside the free text-based search method, MouseBook also allows the user to search for phenotypes associated with data within MouseBook using MP ontology terms (Figure 4). It presents an AJAX-driven expandable tree representation of the MP ontology which allows the user to explore the ontology hierarchy to pick a phenotype term or they can use a phenotype term autosuggest facility to quickly and easily pick a phenotype term to search MouseBook with. When the user clicks on a specific term node of the tree, it then displays the search results comprising any data annotated with that specific term and any subsequent child terms.

MouseBook also aids the user to specifically search using a gene or allele symbol. It utilizes 'auto suggest' technology to facilitate the swift and accurate selection of a gene or allele symbol that is guaranteed to return results from MouseBook. The user can also use MouseBook's advanced stock search, giving the ability to fine-tune their search for a mouse stock by allowing them to select, for example, stocks with specific chromosomes involved, the mutation type involved in the generation of the mouse or which strain background the stock has.

Mousebook enables users to register and login to receive additional functionality such as an update service which will inform them when new data enters the database matching their search, for example a new stock, as well as remembering their shipping details for streamlined ordering. This facility will enable the user to track their orders through MouseBook.

IMPLEMENTATION

MouseBook's web front end is written in PHP utilizing CSS and the JavaScript JQuery framework for improved user interactions. The JavaScript framework enables easy auto suggest functionality as well as a tabbed interface, allowing the user to have quick access to more information without the need to scroll. User registration, requests and search tracking data are stored and encrypted in a MySQL database. MouseBook utilizes Memcache to counter network lag when dealing with web services and reduce database load greatly improving performance.

MouseBook's searchable data sources currently consist of five internal MySQL relational databases containing stock data, imprinting data, chromosomal anomalies data, phenotype annotations and publications with external access to EMMA data via web services. MouseBook's search module utilizes specific 'overview' snapshot database tables created within these databases. This adds the flexibility to use a 'one search method fits all' approach across all datasets as well as not requiring the databases to be in a specific schema. This allows MouseBook instant 'plug and play' functionality to integrate any other MySQL data sources.

Alongside this approach, the MouseBook search module can be integrated with web services (SOAP/REST), for example, a BioMart or external WSDL file. Together, these technologies allow the simple and easy integration of additional data sources into MouseBook, creating catalogs with original data sources ranging from Excel spreadsheets through to datasets with online programmatic access.

The Stocklist Information Management system consists of a MySQL relational database holding the public and private stock information and a reference database built from the marker/gene and allele data provided by

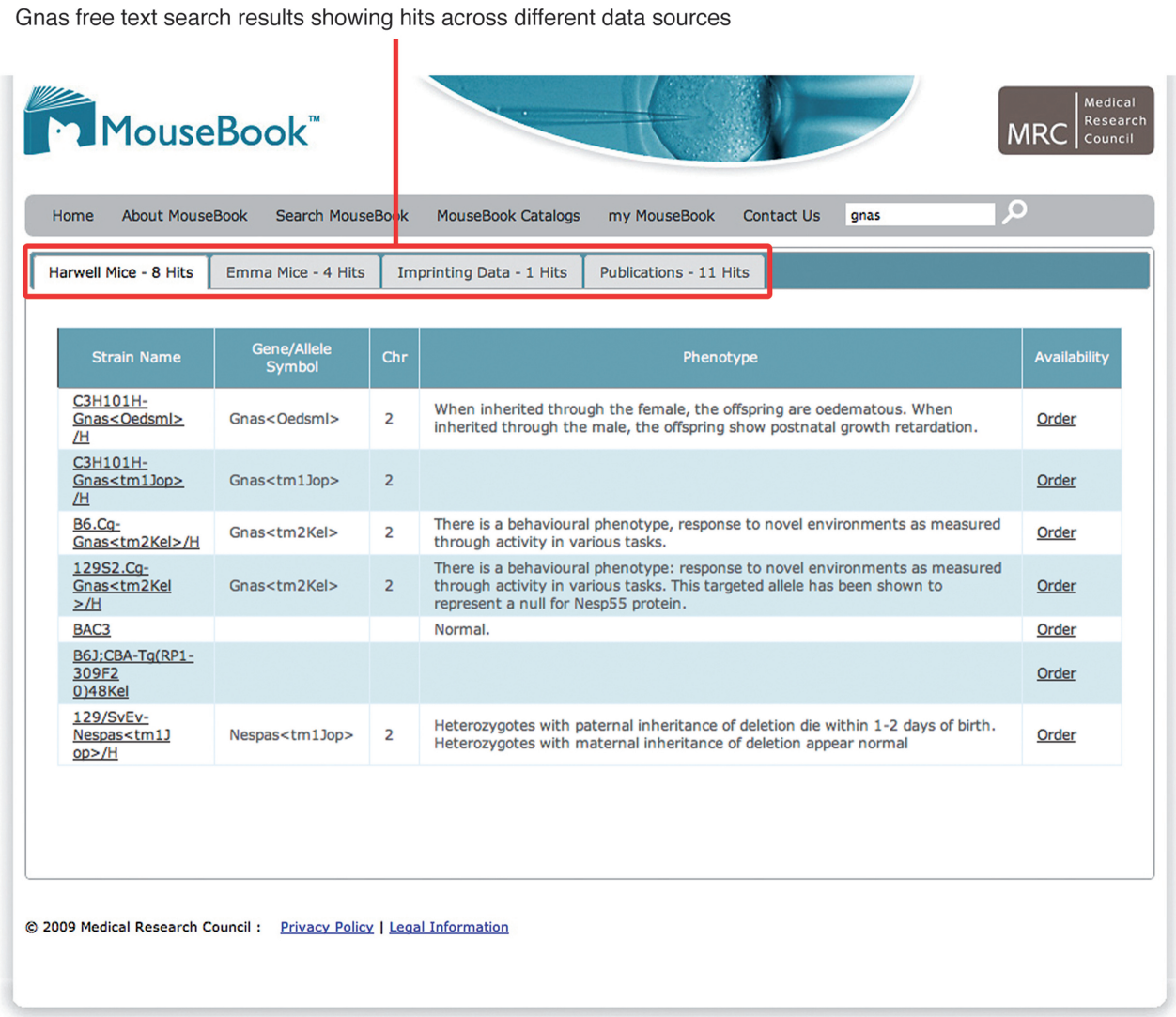

Figure 3. The results of a MouseBook search for the free text 'gnas', a well-known imprinted gene. It shows hits within Harwell Mice, EMMA Mice, Imprinting Data and Publication catalogs having used both the free text 'gnas' (20) as well as the MGI identifier (MGI:95777), automatically ascertained by the search module, to query all MouseBook data sources. In case of the Harwell Mice search result's tab, it shows the Strain Name, involved gene/allele symbols, the chromosome involved, phenotypic description and the ability to order material for all matched stocks.

the MGD public FTP site. An in-house middleware layer then presents the data as a Java object model, and the system is accessed via a Java/Swing graphical user interface.

OATH is a central repository for storing and curating phenotypic annotations either by hand, computationally or via parsing of free-text phenotypic descriptions. Those annotations can be linked to any data point within MouseBook thus allowing the user to quickly and easily search through all of MouseBook's data sources with a phenotype term. Annotations and curator details are stored encrypted in a MySQL relational database with the front end using PHP, JavaScript and CSS. Ontologies are loaded via web services from the Ontology Lookup System (OLS) (19).

MouseBook is an open source project and all source code can be obtained by contacting the authors. Future plans for MouseBook are to make the data as accessible as possible by providing downloads as well as programmatic access.

FUTURE DIRECTIONS

The impact of high-throughput projects generating and characterizing mouse mutants in conjunction with the rapid increase in databases is beginning to be felt by individual researchers trying to identify new mouse models. MouseBook has utilized new approaches in data integration to provide a user-friendly portal that enables users access to a wealth of integrated data. MouseBook's future challenge is to integrate new mouse resource catalogs either from large consortia (e.g. IKMC) or smaller research laboratories into the MouseBook architecture. MouseBook would, therefore, like to outreach to smaller laboratories that may not have the IT infrastructure or expertise in curation to invite them to

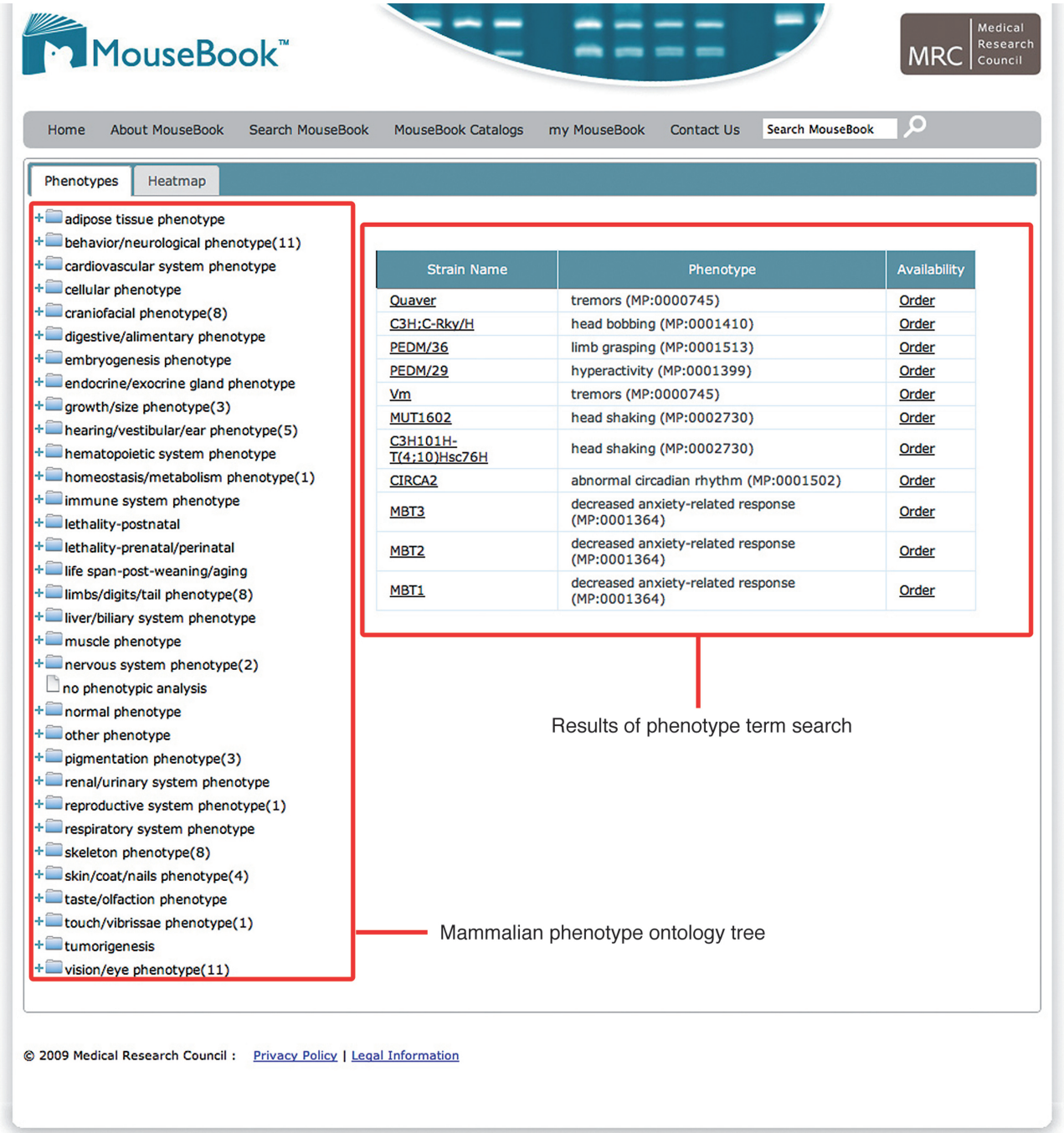

Figure 4. The results of a MouseBook phenotype ontology term search using the high level 'Behaviour/Neurological Phenotype' term (MP:0005286). This hits annotations matching that MP term and any of its child terms. The results are shown displaying the Strain Name, the actual phenotype term that is annotated and the ability to order cryopreserved material.

contact the MouseBook team (info@mousebook.org) who would be happy to provide the information systems to enable their data to be searched or in turn import their data into a MouseBook data source. MouseBook also aims through collaboration with the 'International Committee on Standardized Genetic Nomenclature for Mice' (http://www.informatics.jax.org/nomen) to ensure that new resources integrated in MouseBook are curated and uploaded to IMSR.

Next Generation Sequencing technology will make a significant contribution to the characterization of mutant mouse lines in the next few years, both by allowing sequencing of specific regions to identify SNPs and other mutations and by facilitating other types of characterization such as transcriptomics and ChIP-Seq. It is planned to expand MouseBook to provide access to the results of such analyses being generated at Harwell and elsewhere.

A further challenge for MouseBook is to develop new tools and search mechanisms which will enable users to identify mutants more easily by their phenotype or their similarity to human disease. MouseBook therefore aims to integrate data from high-throughput phenotyping databases such as EuroPhenome with its curated phenotype annotations.

ACKNOWLEDGEMENTS

We acknowledge those researchers and organizations that have provided the data to MouseBook. We thank Damian Smedley and Phil Wilkinson for ensuring data accessibility to the EMMA database and the members of the FESA team at Harwell, who archive and distribute the mouse strains.

FUNDING

UK Medical Research Council. Funding for open access charge: UK Medical Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Rosenthal,N. and Brown,S. (2007) The mouse ascending: perspectives for human-disease models. *Nat. Cell Biol.*, **9**, 993–999.
2. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
3. Brown,S.D., Hancock,J.M. and Gates,H. (2006) Understanding mammalian genetic systems: the challenge of phenotyping in the mouse. *PLoS Genet.*, **2**, e118.
4. Hancock,J.M. and Mallon,A.-M. (2007) Phenobabelomics–mouse phenotype data resources. *Brief. Funct. Genomic. Proteomic.*, **6**, 292–301.
5. Mouse Phenotype Database Integration Consortium. (2007) Integration of mouse phenome data resources. *Mamm. Genome*, **18**, 157–163.
6. Stein,L. (2002) Creating a bioinformatics nation. *Nature*, **417**, 119–120.
7. Stein,L.D. (2008) Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nat. Rev.*, **9**, 678–688.
8. Smedley,D., Swertz,M.A., Wolstencroft,K., Proctor,G., Zouberakis,M., Bard,J., Hancock,J.M. and Schofield,P. (2008) Solutions for data integration in functional genomics: a critical assessment and case study. *Brief. Bioinform.*, **9**, 532–544.
9. Eppig,J.T. and Strivens,M. (1999) Finding a mouse: the International Mouse Strain Resource (IMSR). *Trends Genet.*, **15**, 81–82.
10. Collins,F.S., Rossant,J. and Wurst,W. (2007) A mouse for all reasons. *Cell*, **128**, 9–13.
11. Nuffield Council on Bioethics. (2005) *The Ethics of Research Involving Animals*. Nuffield Council on Bioethics, London.
12. Hagn,M., Marschall,S. and Hrabe de Angelis,M. (2007) EMMA–the European mouse mutant archive. *Brief. Funct. Genomic. Proteomic.*, **6**, 186–192.
13. Blake,J.A., Bult,C.J., Eppig,J.T., Kadin,J.A. and Richardson,J.E. (2009) The Mouse Genome Database genotypes::phenotypes. *Nucleic Acids Res.*, **37**, D712–D719.
14. Smith,C.L., Goldsmith,C.A. and Eppig,J.T. (2005) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.*, **6**, R7.
15. Mallon,A.M., Blake,A. and Hancock,J.M. (2008) EuroPhenome and EMPReSS: online mouse phenotyping resource. *Nucleic Acids Res.*, **36**, D715–D718.
16. Gkoutos,G.V., Green,E.C., Mallon,A.M., Hancock,J.M. and Davidson,D. (2005) Using ontologies to describe mouse phenotypes. *Genome Biol.*, **6**, R8.
17. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
18. Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). http://www.ncbi.nlm.nih.gov/omim/.
19. Cote,R.G., Jones,P., Martens,L., Apweiler,R. and Hermjakob,H. (2008) The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res.*, **36**, W372–W376.
20. Peters,J., Holmes,R., Monk,D., Beechey,C.V., Moore,G.E. and Williamson,C.M. (2006) Imprinting control within the compact Gnas locus. *Cytogenet. Genome Res.*, **113**, 194–201.