D568–D571 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh069

The mouse Gene Expression Database (GXD): updates and enhancements

David P. Hill, Dale A. Begley, Jacqueline H. Finger, Terry F. Hayamizu, Ingeborg J. McCright, Constance M. Smith, Jon S. Beal, Lori E. Corbani, Judith A. Blake, Janan T. Eppig, James A. Kadin, Joel E. Richardson and Martin Ringwald*

The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA

Received September 15, 2003; Accepted September 30, 2003

ABSTRACT

The Gene Expression Database (GXD) is a community resource for gene expression information in the laboratory mouse. By collecting and integrating different types of expression data, GXD provides information about expression profiles in different mouse strains and mutants. Participation in the Gene Ontology (GO) project classifies genes and gene products with regard to molecular functions, biological processes, and cellular components. Integration with other Mouse Genome Informatics (MGI) databases places the gene expression information in the context of mouse genetic, genomic and phenotypic information. The integration of these types of information enables valuable insights into the molecular biology that underlies development and disease. The utility of GXD has been improved by the daily addition of new data and through the implementation of new query and display features. These improvements make it easier for users to interrogate and visualize expression data in the context of their specific needs. GXD is accessible through the MGI website at http://www.informatics.jax.org/ or directly at http://www.informatics.jax.org/menus/expression_menu.shtml.

INTRODUCTION

The completion of the genome sequence of the laboratory mouse has solidified its importance as a model organism for the study of mammalian biology and human disease (1). To learn more about the function of genes in the mammalian genome, information about where and when the genes are expressed becomes critical for understanding the molecular mechanisms that underlie development and disease. The laboratory mouse is particularly tractable for analysis of gene expression since tissues from all stages of development and many different strains and mutants are readily available. The Gene Expression Database (GXD) has been designed as an open-ended system that can integrate many types of gene-expression data with the biology and genetics of the laboratory mouse (2–7). GXD provides information about both mRNA and protein expression by capturing details of expression from a variety of assays such as RNA *in situ* hybridization, immunohistochemistry, northern blot, western blot, RT–PCR and cDNA source data. Expression information is standardized by using a hierarchical mouse anatomical dictionary constructed in collaboration with our Edinburgh colleagues (8), by employing a set of controlled vocabularies describing pattern and level of expression, and by using standard gene, strain and allele nomenclature for recording the ancillary experimental data. The standardized expression information is supported by digitized images of original data linked to the expression records. Through participation in the Gene Ontology (GO) project GXD incorporates functional information about gene products with expression information (9). GXD is fully integrated with the other databases of the Mouse Genome Informatics (MGI) resource (10,11). The MGI resource also provides comprehensive links to external resources such as sequence databases, OMIM, MEDLINE and databases from other model species (12–21). The integration of these large amounts of information puts expression data annotated in GXD in a much larger biological and analytical context.

GXD is implemented in the Sybase relational database management system. The database is continuously curated and updates are made available on a daily basis. Users access data primarily by web-based query forms, but direct SQL access to the system is also available via our user support group. GXD and its query interfaces have been described in detail previously (5,7). Here we report recent progress on data content, query capabilities and results visualization.

DATA CONTENT

The Gene Expression Literature Index

The Gene Expression Literature Index is a tool to rapidly find publications with specific types of expression information using a variety of parameters. References are indexed with respect to specific genes, developmental stages, expression assays and bibliographic information. In addition, we provide the capability to search for text strings that appear in

*To whom correspondence should be addressed. Tel: +1 207 288 6436; Fax: +1 207 288 6132; Email: ringwald@informatics.jax.org publication abstracts. All journal articles containing data on endogenous gene expression during mouse development from 1993 to the present and major developmental journals from 1990 are indexed. Recently, we have also begun to index information from publications that use 'knock-in' reporter genes to study gene expression. We continue to keep the Gene Expression Literature Index up to date. As of September 15, 2003, the index includes 31 917 entries covering 8842 references and expression information for 5560 genes.

Gene expression data

GXD includes data from a variety of assays including RNA *in situ* hybridization, immunohistochemistry, northern and western blot, RNase protection and RT–PCR experiments. So far, the primary source of these types of expression data in GXD is hand annotation of manuscripts by GXD curators. The data are extracted from the literature and entered into GXD on a daily basis. Curators enter data related to individual assays that are defined as the analysis of expression of one gene in one or multiple samples by a specific method using a specific probe under specific experimental conditions.

We also have begun acquiring large sets of expression data from large-scale RNA *in situ* hybridization screens. In direct collaboration with respective laboratories, a combination of hand annotation and bulk data downloading is used to capture large sets of detailed expression data. An example of this type of data load is from publications describing the expression of mouse genes that are orthologous to genes on human chromosome 21 (22,23). These reports contain large amounts of expression information as Supplementary Material that has now been fully integrated into GXD.

GXD also captures expression data that have been analyzed in mutant mice. More than half of the assays in GXD include expression analysis in mutants. As of September 15, 2003, GXD includes 113 862 results from 9561 assays covering expression information for 2827 genes. A significant part of these data represent complex RNA *in situ* and immunohistochemistry expression patterns. Most of these data are linked to images from primary expression data. Recently we began capturing data from reporter gene 'knock-in' experiments. 'Knock in' experiments use homologous recombination to place reporter genes into the genome under the control of endogenous gene regulatory elements. These reporter genes are then used as sensitive indicators of gene expression. 'Knock in' data are captured with respect to the endogenous gene they are used to study and the detection method used to identify the reporter gene: RNA *in situ* hybridization, immunohistochemistry or direct assay. As the amount and type of data captured by GXD increases, GXD provides a more and more complete representation of gene expression patterns in the mouse.

In addition to hand-curated data, GXD also represents expression information from cDNAs and their source tissues that are obtained in large data downloads. cDNA source information is available from clones available from the IMAGE Consortium, the Riken FANTOM project and the NIA clone sets. In particular, we collaborated with our colleagues from the Mouse Genome Database (MGD), the Mouse Genome Sequence Project and other members of the FANTOM consortium to annotate 60 770 fully sequenced mouse cDNAs (24). By using a combination of manual curation and computational analysis we were able to group these cDNAs into 33 409 clustered transcription units, to partially characterize and classify these transcription units with regard to GO terms, and to map many of them to the mouse genome sequence (11,24,25). The cDNA source data from these clusters of cDNAs were loaded into GXD and provide an integrated view of the expression patterns of these genes. As of September 15, 2003, GXD contained data for 427 548 mouse cDNA clones. These clones represent 19 002 markers and their library sources represent 333 different tissues. The integration of cDNA data, the many newly established links between cDNAs and genes, and the classification of these cDNAs and genes according to GO terms provide an important basis for the incorporation and querying of microarray-based expression data in GXD.

Anatomy

Developmental expression patterns in GXD are described using an anatomical dictionary for mouse development that has been developed by our colleagues from the Edinburgh Mouse Atlas project (8). The anatomical dictionary is broken down into the 26 Theiler stages for mouse development (26) and lists anatomical terms for each developmental stage in a hierarchical tree. As part of the GXD project, we have recently developed an extensive anatomical dictionary for the adult mouse (http://www.informatics.jax.org/searches/AMA_form.shtml). This ontology is structured as a directed acyclic graph, in which an anatomical term can be represented as a child of more than one hierarchical parent term using 'part of' and 'is a' relationships. For example, the heart is represented as a part of the cardiovascular system and as a type of ('is a') thoracic cavity organ. Thus, the anatomical dictionary for the adult mouse structures the anatomy both spatially and functionally. Together, the anatomical ontologies for the developmental and adult mouse allow for a robust spatial description of gene activity throughout the life of the laboratory mouse. In addition, the vocabularies can be shared by annotators to describe other areas of mouse biology, for example they can be used in combination with other vocabularies to describe tissue-specific biological processes or to describe the phenotypic effects of a mutation on an anatomical structure (27).

WEB REPRESENTATION

Query forms

The Gene Expression Data query form (http://www.informatics.jax.org/searches/expression_form.shtml) provides access to data from RNA *in situ* hybridization, immunohistochemistry, northern blot, western blot, RT–PCR and RNase protection experiments. The basic query form has been described in detail previously (7) and can be used to ask complex queries using a variety of search parameters including genes, chromosomal location, tissues, stages of development and the gene function. Recently, we have enhanced the existing 'Gene Expression Data Query Form' and we have added an 'Expanded Gene Expression Data Query Form' to permit more sophisticated searches. We have added the ability to search for expression information in mutant animals. For example, it is now possible to construct a query such as: 'Show me what the expression of *Nef3* looks like in animals that are mutant for the *Hoxa3* gene' or 'What data is available describing the expression of genes in the eyes of *Pax6* mutants?' The new 'Expanded Gene Expression Data Query Form' allows users to formulate Boolean expressions when querying for anatomical structures. This new capability permits queries such as: 'Show me all genes that are expressed in the brain and the limb' or 'Show me the genes that are expressed in the hindbrain, but not in the midbrain'. These types of queries enhance the ability to interrogate gene expression patterns from an anatomical perspective.

Another way to obtain data from an anatomical perspective is to use the Anatomical Dictionary Browser (http://www.informatics.jax.org/searches/anatdict_form.shtml). This browser permits users to search for and locate specific anatomical terms in the dictionary and to scan the anatomical dictionary structure by clicking on terms to expand and contract the vocabulary. For each anatomical structure, a link is provided that will return all of the expression results associated with the structure or its substructures. We have added a feature allowing a refined query where specific expression results associated with an anatomical structure can be further restricted using criteria found on the regular Gene Expression Data Query Form.

Results visualization

Several enhancements have been made to the visualization of expression data in the MGI database system. For each marker in MGI, a Gene Detail page is available that summarizes the knowledge about the gene. We have added an expression summary to the Gene Detail page that lists the Theiler stages where the gene has been analyzed and gives a summary of the types of expression assays that have been used to study the gene. In addition the number of tissues where gene expression has been studied is displayed. The number is hot-linked to a tissue summary page listing the tissues and results where the gene was found to be expressed or not to be expressed in that tissue. This page in turn links to expression details about the experiment.

We have also made improvements to the results visualization for many of the GXD Query Form returns. The results from the GXD index have been completely reformatted. Results are now reported in tabular format indicating the number of references associated with expression data for a given assay at a given stage of development. The reference numbers are hot-linked to a reference summary page where each reference can be inspected independently for both index information and full expression data information. Significant improvements have also been made to the gene expression assay results summary page. If results are associated with mutant alleles, then they can be easily identified and the mutant genotype is listed in the summary table. Results are also now sorted according to the anatomical hierarchy rather than alphabetically. This makes scanning the results much simpler because, for example, the brain and all of its substructures for a given Theiler stage are now grouped together. For example, brain is now grouped close to forebrain rather than branchial arch.

THE GENE EXPRESSION NOTEBOOK

GXD has developed The Gene Expression Notebook (GEN), a tool that can be used as a laboratory notebook to store expression data and for electronic submission of expression data to GXD. GEN is implemented in Microsoft Excel and details about it have been described previously (6). GEN has now been expanded and can store all types of assays currently stored in GXD. The notebook is available at: http://www.informatics.jax.org/mgihome/GXD/GEN/. We welcome feedback and data submissions. Data submissions will be reviewed by GXD curators and will receive accession numbers that can be cited in publications.

FUTURE DIRECTIONS

GXD will continue to acquire gene expression data through curation of the literature and electronic data submission and to work with large-scale data providers in incorporating their data using bulk-load procedures. In the near future, GXD will be expanded to include microarray expression data. Our work will focus on adding value to array data by integrating them with other types of expression data, and with genomic, genetic and phenotype data for the laboratory mouse. Our Edinburgh collaborators have already begun to map RNA *in situ* expression data from GXD into the 3D Atlas/graphical gene expression database for mouse development [EMAGE, http://genex.hgu.mrc.ac.uk/Emage/database/intro.html (28)]. We will continue to integrate GXD and EMAGE to generate the Mouse Gene Expression Information Resource, which will fully combine text-based and graphical means for storing and analyzing expression data (2). GXD captures information about probes used in each assay for gene expression. As better sequence representation is achieved by MGI, we will be able to better correlate probe information with other genomic information such as alternative exon usage and protein isoforms. We will continue to develop the integration between expression data and genetic and phenotypic data represented in the MGD by cultivating refined links between expression and genotype and by sharing vocabularies such as the anatomical dictionary to describe expression and phenotype data.

USER SUPPORT

GXD provides support to its users through online documentation and a dedicated User Support staff. User support can be contacted by email (mgi-help@informatics.jax.org), by fax (+1 207 288 6132) or by telephone (+1 207 288 6445).

CITING GXD

To reference the database itself, please cite this article. For referring to specific GXD data, we suggest the following format: 'These data were retrieved from the Gene Expression Database (GXD), Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine, USA, World Wide Web (http://www.informatics.jax.org)'. [Type in date (month, year) when you retrieved the data cited.]

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank our colleagues from the other MGI projects for their contributions to the GXD project and to the larger integrated MGI resource, and Drs Tim O'Brien and Joel Graber for critically reading the manuscript. GXD is supported by NIH grant HD33745. The GO project is supported by NIH grant HG02273.

REFERENCES

1. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
2. Ringwald,M., Baldock,R., Bard,J., Kaufman,M., Eppig,J.T., Richardson,J.E., Nadeau,J.H. and Davidson,D. (1994) A database for mouse development. *Science*, **265**, 2033–2034.
3. Ringwald,M., Davis,G.L., Smith,A.G., Trepanier,L.E., Begley,D.A., Richardson,J.E. and Eppig,J.T. (1997) The Mouse Gene Expression Database GXD. *Semin. Cell Dev. Biol.*, **8**, 489–497.
4. Ringwald,M., Mangan,M.E., Eppig,J.T., Kadin,J.A., Richardson,J.E. and the Gene Expression Database Group. (1999) GXD: a Gene Expression Database for the laboratory mouse. *Nucleic Acids Res.*, **27**, 106–112.
5. Ringwald,M., Eppig,J.T., Kadin,J.A., Richardson,J.E. and the Gene Expression Database Group (2000) GXD: a Gene Expression Database for the laboratory mouse: current status and recent enhancements. *Nucleic Acids Res.*, **28**, 115–119.
6. Begley,D.A. and Ringwald,M. (2002) Electronic tools to manage gene expression data. *Trends Genet.*, **18**, 108–110.
7. 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.
8. Bard,J.B.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.
9. Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
10. 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.
11. Bult,C.J., Blake,J.A., Richardson,J.E., Kadin,J.A., Eppig,J.T. and the Mouse Genome Database Group (2004) The Mouse Genome Database (MGD): integrating biology with the genome. *Nucleic Acids Res.*, **32**, D476–D481.
12. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
13. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
14. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
15. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D.R. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
16. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
17. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
18. Quackenbush,J., Cho,J., Lee,D., Liang,F., Holt,I., Karamycheva,S., Parvizi,B., Pertea,G., Sultana,R. and White,J. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.*, **29**, 159–164.
19. Twigger,S., Lu,J., Shimoyama,M., Chen,D., Pasko,D., Long,H., Ginster,J., Chen,C.F., Nigam,R., Kwitek,A. *et al.* (2002) Rat Genome Database (RGD): mapping disease onto the genome. *Nucleic Acids Res.*, **30**, 125–128.
20. Gas,S., Eggen,A., Samson,F., Christophe,C., Mungall,C., Bessieres,P. and Leveziel,H. (1996) The BOVMAP database. XXVth International Conference on Animal Genetics, 21–25 July 1996, Tours, France. *Animal Genet.*, **27** (Suppl. 2), 59.
21. Hu,J., Mungall,C., Law,A., Papworth,R., Nelson,J.P., Brown,A., Simpson,I., Leckie,S., Burt,D.W., Hillyard,A.L. *et al.* (2003) The ARKdb: genome databases for farmed and other animals. *Nucleic Acids Res.*, **29**, 106–110.
22. Reymond,A., Marigo,V., Yaylaoglu,M.B., Leoni,A., Ucla,C., Scamuffa,N., Caccioppoli,C., Dermitzakis,E.T., Lyle,R., Banfi,S. *et al.* (2002) Human chromosome 21 gene expression atlas in the mouse. *Nature*, **420**, 582–586.
23. Gitton,Y., Dahmane,N., Baik,S., Ruiz,I., Altaba,A., Neidhardt,L., Scholze,M., Herrmann,B.G., Kahlem,P., Benkahla,A. *et al.* (2002) A gene expression map of human chromosome 21 orthologues in the mouse. *Nature*, **420**, 586–590.
24. Baldarelli,R.M., Hill,D.P., Blake,J.A., Adachi,J., Furuno,M., Bradt,D., Corbani,L.E., Cousins,S., Frazer,K.S., Qi,D. *et al.* (2003) Connecting sequence and biology in the laboratory mouse. *Genome Res.*, **13**, 1505–1519.
25. The FANTOM Consortium and The RIKEN Genome Exploration Research Group Phase I & II Team (2002) Analysis of the mouse transcriptome based on functional annotation of 60 770 full-length cDNAs. *Nature*, **420**, 563–573.
26. Theiler,K. (1989) *The House Mouse Atlas of Embryonic Development*. Springer-Verlag, New York, NY.
27. 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.
28. Davidson,D.R., Bard,J.B.L., Kaufman,M.H. and Baldock,R.A. (2002) The Mouse Atlas Database: a community resource for mouse development. *Trends Genet.*, **17**, 49–51.