The mouse Gene Expression Database (GXD): 2017 update

Jacqueline H. Finger, Constance M. Smith, Terry F. Hayamizu, Ingeborg J. McCright, Jingxia Xu, Meiyee Law, David R. Shaw, Richard M. Baldarelli, Jon S. Beal, Olin Blodgett, Jeff W. Campbell, Lori E. Corbani, Jill R. Lewis, Kim L. Forthofer, Pete J. Frost, Sharon C. Giannatto, Lucie N. Hutchins, Dave B. Miers, Howie Motenko, Kevin R. Stone, Janan T. Eppig, James A. Kadin, Joel E. Richardson and Martin Ringwald[*]

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

ABSTRACT

The Gene Expression Database (GXD; www.informatics.jax.org/expression.shtml) is an extensive and well-curated community resource of mouse developmental expression information. Through curation of the scientific literature and by collaborations with large-scale expression projects, GXD collects and integrates data from RNA *in situ* hybridization, immunohistochemistry, RT-PCR, northern blot and western blot experiments. Expression data from both wild-type and mutant mice are included. The expression data are combined with genetic and phenotypic data in Mouse Genome Informatics (MGI) and made readily accessible to many types of database searches. At present, GXD includes over 1.5 million expression results and more than 300 000 images, all annotated with detailed and standardized metadata. Since our last report in 2014, we have added a large amount of data, we have enhanced data and database infrastructure, and we have implemented many new search and display features. Interface enhancements include: a new Mouse Developmental Anatomy Browser; interactive tissue-by-developmental stage and tissue-by-gene matrix views; capabilities to filter and sort expression data summaries; a batch search utility; gene-based expression overviews; and links to expression data from other species.

INTRODUCTION

Developmental gene expression information from wild-type and mutant mice provides crucial insights into the molecular mechanisms of human development, differentiation and disease. The longstanding objective of GXD has been to capture and integrate these complex and heterogeneous expression data, generated by many laboratories and scattered through thousands of publications, and to make them readily accessible to researchers via biologically and biomedically relevant searches. Designed from the beginning as an open-ended system that can robustly integrate different types of expression data, GXD collects RNA and protein expression information from RNA *in situ* hybridization, immunohistochemistry, *in situ* reporter (knock in), RT-PCR, northern blot, and western blot experiments (1,2). Data are acquired from the literature and through collaboration with projects that generate the types of data GXD collects at a large scale. All these data are reviewed and annotated by GXD Curators in standardized ways by making extensive use of controlled vocabularies and ontologies. As an integral component of the larger Mouse Genome Informatics (MGI) resource (3–5), GXD combines the expression data with genetic, functional, phenotypic, and disease-oriented data, thereby enabling unique and powerful search capabilities. GXD and its user interfaces have been described previously (6,7). Here, we focus on progress made since our last report in the *NAR* Database Issue (8).

DATA CONTENT AND PROGRESS IN DATA ACQUISITION

Comprehensive literature survey

We systematically survey journals to find all publications examining endogenous gene expression during mouse development. In a first curation step for each paper, we annotate the genes and ages analyzed and the expression assay types used. Annotations are based on the entire publication, including supplemental information, and employ official nomenclature for genes. This information, combined with bibliographic information from PubMed, is made available for searches via the Gene Expression Literature Query form (http://www.informatics.jax.org/gxdlit). GXD's liter-

---

[*]To whom correspondence should be addressed. Tel: +1 207 288 6436; Fax: +1 207 288 6830; Email: martin.ringwald@jax.org

Figure 1. The Assay Details Page—annotated expression data are displayed with citation information, gene, and experimental information, such as the assay type, probe visualization method and link to probe specifics. The results section of the page includes the level and pattern of expression in anatomical structures (annotated using the Mouse Developmental Anatomy Ontology). Image panes accompany the result annotations when available. The pane can be viewed in the context of the full figure by clicking on the image or the link in the result table. Specimen information includes the age, genetic background and specimen preparation methods. Further details can be viewed by expanding the 'more' toggle (boxed region indicated by red circle & arrow). If the specimen is from a mutant mouse, the official allele name is shown with links to all the information for that allele in MGI, including all the other expression data for this specific mouse mutant.

ature content records are comprehensive and up-to-date from 1990 to the present. GXD has records for over 24 300 references and >15 500 genes. The Gene Expression Literature Query provides researchers with an effective tool for finding publications with specific expression data, and it helps GXD Curators to prioritize papers for detailed expression annotation.

Detailed expression data

In a second curation step, the expression data are annotated in detail. For each expression assay, we record information about the gene studied, the strength and pattern of expression in specific anatomical structures, the probes and experimental conditions used, and the age and genetic background of the specimen(s) analyzed. Images of the data accompany the annotations when available. Standard gene,

Figure 2. The Mouse Developmental Anatomy Browser—divided into three sections, the Anatomy Search, the Anatomical Term Detail and the Anatomical Tree View, the user can quickly find terms and explore the mouse anatomy ontology. As indicated in the Term Detail section, individual structures can have multiple parents; 'branchial arch endoderm' has Parent Terms 'embryo endoderm' and 'branchial arch'). The EMAPA anatomy ontology terms are represented in the time-independent 'abstract' version (as evidenced by the 'Present at' range of Theiler stages). Individual Theiler stages can be selected to view an anatomical term in the context of a time-dependent hierarchy (boxed region, blue arrow). In the Anatomical Tree View the chosen term is highlighted and toggles allow the user to explore up and down the tree. A link to the expression results summaries for the selected anatomical structure is provided (red arrow).

mouse strain and allele nomenclature, controlled vocabularies and an anatomy ontology are employed to enable thorough data integration and search capabilities (Figure 1). Since our last report, we have continued to curate and integrate expression data from the literature and from large-scale expression projects, including new RNA *in situ* data from the GUDMAP project (9) and a first set of *in situ* reporter (knock- in) data from the International Mouse Phenotype Consortium (IMPC; 10). As of September 2016, GXD contains detailed expression data for ~14 300 genes,

including data from numerous strains of wild-type mice and from >3100 mouse mutants. GXD now holds >300 000 images and >1.5 million expression result annotations.

ENHANCEMENTS OF DATA STRUCTURE AND WEB INTERFACE

We have migrated our system to utilize an enhanced version of the mouse anatomical ontology, a key component of GXD. Features added to GXD's website include:
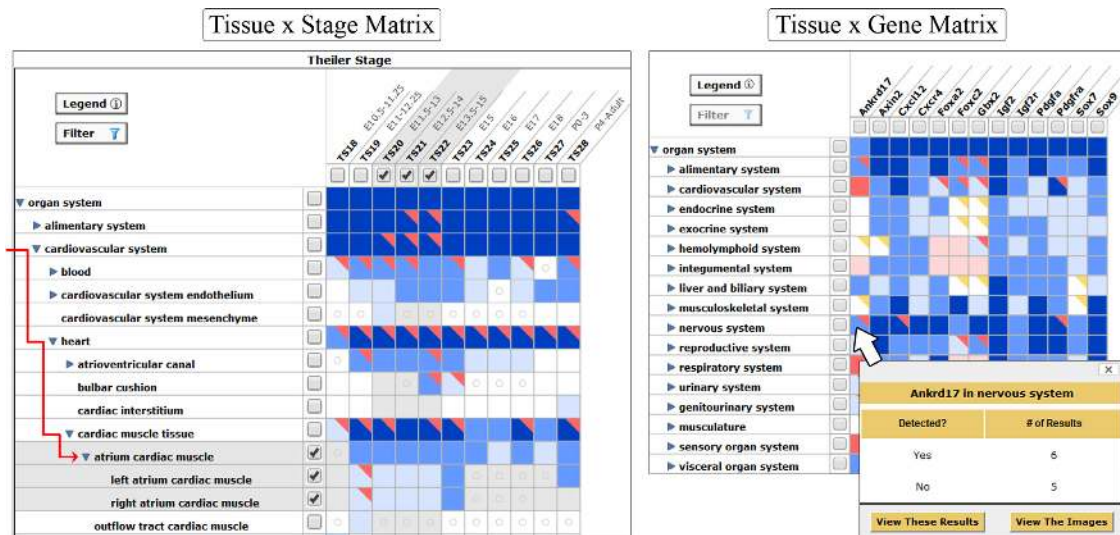
Figure 3. Interactive matrix views—the Tissue-by-Stage (left) and the Tissue-by-Gene (right) matrices are interactive overviews of expression data that can be expanded along the tissue axis (see red arrow) to view more detailed anatomical structures and associated expression data. The Legend, accessed using the top left button, explains the colors and symbols. Clicking an individual cell in either matrix view opens a summary window of the data represented in that cell (shown in the Tissue-by-Gene Matrix by the white arrow). To filter, select individual rows and columns (indicated by the checkmarks and highlighting on the Tissue-by-Stage Matrix) and use the Filter button located at the top left.

a new Mouse Developmental Anatomy Browser; interactive tissue-by-developmental stage and tissue-by-gene matrix views; capabilities to filter and sort expression data summaries; a batch search utility; gene-based expression overviews; and links to expression data from other species. We now describe these new components in greater detail and then illustrate how they have been combined into a unified web interface.

Enhanced mouse anatomy ontology and new mouse developmental anatomy browser

In GXD, time and space of gene expression is recorded using the mouse anatomical ontology. The original version of the mouse developmental anatomy ontology listed the anatomical structures for each developmental (Theiler) stage in separate uni-parental hierarchies. While this ontology served GXD well for many years, it did have shortcomings. The hierarchy for each Theiler stage had to be maintained separately, making updates labor intensive and complicated. Further, because anatomical structures could have only one hierarchical parent, they could not be represented, and searched for, as part of different anatomical systems. For example, if the brain was represented as part of the nervous system, it could not be part of the head. Together with our colleagues from the Edinburgh Mouse Atlas Project (EMAP), we therefore developed a new 'abstract' (non-timed) version of the anatomy ontology; we refer to this as EMAPA (11,12). In this representation, anatomical structures are listed with the developmental stage range over which they exist. Thus, only the 'abstract' ontology must be maintained and, based on the stage-range information, stage-specific ontologies can be instantiated. Further, the new version of the anatomy ontology is structured as a multi-parental hierarchy (directed acyclic graph). We have now migrated GXD to use this new mouse development

anatomy ontology, and we have developed a new Mouse Developmental Anatomy Browser (http://www.informatics.jax.org/vocab/gxd/anatomy). This browser allows users to search for and locate specific anatomical structures, to navigate through the abstract and stage-specific anatomical hierarchies, to explore the anatomy from multiple perspectives, and to obtain associated expression data (Figure 2).

Interactive matrix views of expression data

We have developed two interactive matrix views that provide intuitive high-level summaries of expression data from which one can easily move to views of greater detail (Figure 3). The tissue-by-developmental stage matrix presents a global overview of the spatio-temporal expression patterns of genes. The tissue-by-gene matrix enables a comparison of expression patterns. The anatomy axis of both types of matrices can be expanded to view more detailed structures based on the hierarchical organization of the anatomy. Further, matrix rows and columns can be selected to refine the data set.

Filtering and interactivity of gene expression data summaries

Each expression data search in GXD now yields a page with six tabbed data summaries; one each for genes, assays, assay results and images that match the search parameters, as well as the newly added matrix views described above (Figure 4). Users can easily switch between the different views. We have implemented utilities to filter these summaries by anatomical system, assay type, developmental stage, detection of expression (yes or no), and wild-type or mutant specimen. These filters can be applied in any view, and resulting changes of the data set are propagated to the other summaries.
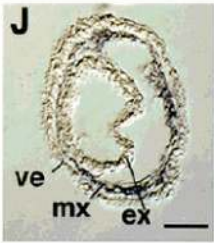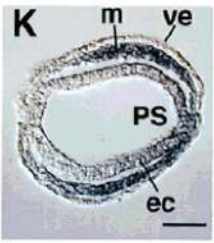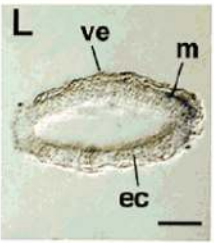
Figure 4. GXD expression data summaries—six views (indicated by red arrow) are available for expression search results: Genes, Assays, Assay Results, Images (shown; tab indicated by blue box), Tissue x Stage Matrix and Tissue × Gene Matrix. Refinements to the data set can be made using the filters (red box), including Anatomical System, Assay Type, Detected? (yes or no), Theiler Stage and Wild type? (or mutant). When applied, these filters (as well as row/column filters on the matrix views) are propagated to all of the data summaries. Sort options are available for many of the columns in the summaries. The red circle indicates the new Specimen Type sort available on the image summary, where images can be sorted into whole mount or section, etc. The Genes Summary and the Assay Results Summary have download or export options for further analysis of the expression data. The data summaries shown here have resulted from the query: 'Which genes located in the genomic region chr6:86000000-142000000 are expressed in the germ layer at E7-8.5 of mouse development?', thus illustrating the use case of a candidate gene search.

Sorting on expression data summaries and exporting data to other applications

The tabular expression data summaries can be sorted by a variety of parameters (e.g. age, gene, structure). A newly added option is the capability to sort the image summary based on specimen type (section, whole mount, optical section, or section from whole mount) (Figure 4). The Genes and Assay Results summary can be exported as text or Excel™ files. We have recently added the capability to export the Genes summary directly into MouseMine, which is discussed below. Thus, gene sets matching specified search criteria in GXD can be exported for further analysis.

New search capabilities

The GXD Home Page provides different search tools to access the detailed expression data: the Gene Expression Data Query; the Mouse Developmental Anatomy Browser (described above); and the Gene Expression Data Matrix that shows all of GXD's 1.5 million expression results in the tissue-by-developmental stage matrix view and can be used for interactive data exploration. The most versatile search utility of these is the Gene Expression Data Query (http://www.informatics.jax.org/gxd). It has three tabbed Search modes: Standard Search, Differential Expression Search, and Batch Search. The Standard Search allows users to query by many different parameters. We have added the capability to search for expression data of genes located within a specific genomic interval, thus facilitating the identifica-
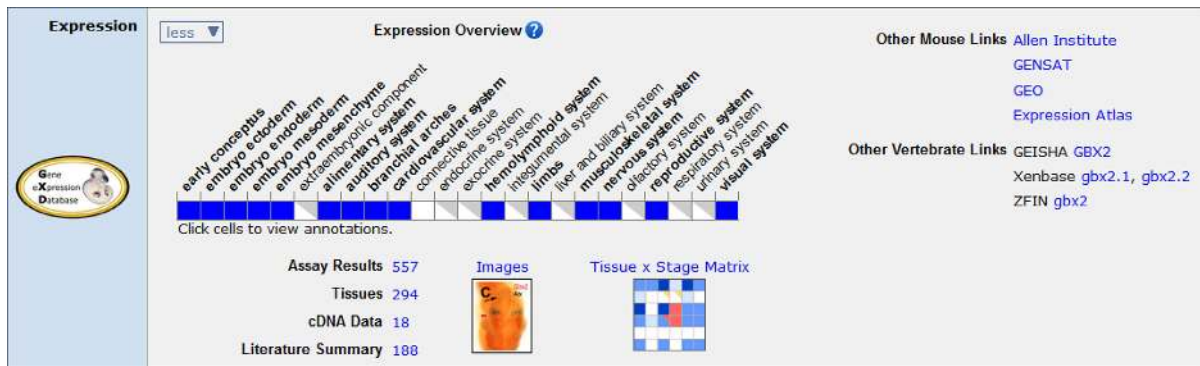
Figure 5. The expression section of the MGI gene detail page—the grid of anatomical terms gives a high-level overview of the expression data for the gene. This figure shows the overview for the gene *Gbx2*. Blue cells indicate that expression was detected in wild type. Gray triangles indicate either absence of expression in wild-type or expression data from mutant specimens. Click in an individual cell in the grid to view the Tissue-by-Stage Matrix for an individual high-level anatomical structure. Click on the matrix icon to view the Tissue-by-Stage Matrix for all high-level anatomical structures. Links to other summaries, such as Images or Assay Results are also included. Links to expression data for the gene at external resources are also available. These include links to other mouse (Allen Institute, GENSAT, GEO, Expression Atlas) and vertebrate (GEISHA, Xenbase, ZFIN) databases.
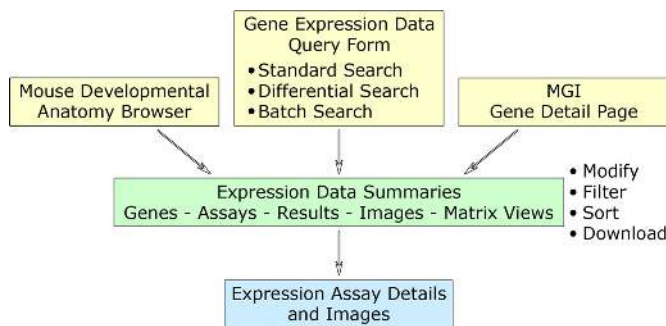
Figure 6. A unified database interface—GXD provides numerous entry points to mouse developmental expression data, such as via the Gene Expression Data Query Forms, the Mouse Developmental Anatomy Browser or the MGI Gene Detail Page. All these entry points lead to the same multi-tabbed summaries. Users are provided with the ability to filter, sort and interactively refine the data, download data for further analysis, and view the details and images from many different perspectives.

tion of candidate genes. The Differential Expression Search allows researchers to query for genes that are expressed in some anatomical structures but not others and/or in some developmental stages but not others. The newly added Batch Search enables users to query by a list of genes. Thus, a researcher who has identified a set of genes of interest, for example through RNA-seq experiments, can quickly find all RNA *in situ* or immunohistochemistry results and images for these genes in GXD.

Gene-based overviews and links to expression data

Gene Detail pages constitute central hubs within MGI because they provide links to all the curated information available for a given gene, as well as links to many external resources. Recent reorganization and enhancements to the gene detail page have created a more graphical and informative view in the gene expression section (Figure 5), as well as in other sections of this page. An overview of a gene's expression is presented via a grid of high-level anatomical structures. Clicking on a cell in this grid will display the anatomical structures' data in a tissue-by-developmental

stage matrix. The gene expression section of the Gene Detail page also provides links to all the expression images, the full Tissue-by-Stage Matrix, and the other expression related summaries for that gene. Gene-based links to expression data from external resources are provided as well. We continue to maintain links to mouse expression data from the Allen Institute (13), GENSAT (14), GEO (15) and EBI's Expression Atlas (16). We have added links to expression data from other vertebrate model organisms highly relevant for developmental research: zebrafish (ZFIN, 17), Xenopus (Xenbase, 18) and chicken (GEISHA, 19).

Combining the web components into a unified database front end

Figure 6 illustrates how the different components described above are connected. The Mouse Developmental Anatomy Browser, the Gene Expression Data Query and the expression section of the Gene Detail page all lead to the multi-tabbed gene expression data summaries. The summaries can be interactively refined through various filtering options; refinements in one summary are propagated to the other summaries. Users can zoom in to the desired sets of data and then follow the links to the detailed expression assay records.

GXD and MouseMine

GXD's expression data are also available through MouseMine. The link (http://www.mousemine.org/mousemine/begin.do#Expression) from the GXD home page leads to the Expression tab of the MouseMine front page. Several 'canned queries' provide a starting point for accessing GXD's expression and anatomy data, and learning the MouseMine conventions. One query combines expression and protein–protein interaction data and lets users search for genes that are expressed in a given anatomical structure and known to interact with a specified gene. This query cannot currently be done in GXD because molecular interaction data is not yet integrated in GXD/MGI. While the GXD interface has been designed

for easy use by biologists and biomedical researchers, the native MouseMine web interface offers computational biologists, as well as others who take the time to become familiar with it, additional flexibility and utilities.

To mention just a few examples: MouseMine offers extensive customized and iterative querying capabilities; gene sets defined by expression criteria can be subjected to enrichment analysis (e.g. for phenotype or human disease correlation); and expression data can be combined with other types of data not yet in GXD or MGI. Further, MouseMine provides comprehensive programmatic access to GXD's data, essential for those wishing to create data pipelines or software tools. Thus, MouseMine complements the utilities provided by GXD. A detailed description of MouseMine is beyond the scope of this article. Readers are referred to MouseMine online documentation and to (20). MouseMine is updated with expression data and other data from MGI on a weekly basis, as indicated in the top right corner of each MouseMine web page.

USER SUPPORT

GXD provides support to its users through a dedicated User Support staff, detailed on-line documentation, and FAQs. User Support can be contacted via email at mgi-help@jax.org or by clicking the User Support link at the bottom of our web pages. The online documentation can be accessed by clicking on the question mark at the top of most pages. FAQs (and other useful links) can be found on the GXD home page (http://www.informatics.jax.org/expression.shtml).

CITING GXD

The following citation format is suggested when referring to data downloaded from GXD: These data were retrieved from the Gene Expression Database (GXD), Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine, USA (URL: http://www.informatics.jax.org) on [type in date (month, year) when you retrieved the data cited]. To reference the database itself, please cite this article.

REFERENCES

1. 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.
2. Ringwald,M., Mangan,M.E., Eppig,J.T., Kadin,J.A. and Richardson,J.E. (1999) GXD: a gene expression database for the laboratory mouse. The Gene Expression Database Group. *Nucleic Acids Res.*, **27**, 106–112.
3. Bult,C.J., Eppig,J.T., Blake,J.A., Kadin,J.A., Richardson,J.E. and the Mouse Genome Database Group. (2016). The mouse genome database 2016. *Nucleic Acids Res.*, **44**, D840–D847.
4. Smith,C.M., Finger,J.H., Hayamizu,T.F., McCright,I.J., Xu,J., Eppig,J.T., Kadin,J.A., Richardson,J.E. and Ringwald,M. (2015) GXD: a community resource of mouse Gene Expression Data. *Mamm. Genome*, **26**, 314–324.
5. Drabkin,H.J., Christie,K.R., Dolan,M.E., Hill,D.P., Ni,L., Sitnikov,D. and Blake,J.A. (2015) Application of comparative biology in GO functional annotation: the mouse model. *Mamm. Genome*, **26**, 574–583.
6. Finger,J.H., Smith,C.M., Hayamizu,T.F., McCright,I.J., Xu,J., Eppig,J.T., Kadin,J.A., Richardson,J.E. and Ringwald,M. (2015) The mouse gene expression database: New features and how to use them effectively. *Genesis*, **53**, 510–522.
7. Smith,C.M., Finger,J.H., Kadin,J.A., Richardson,J.E. and Ringwald,M. (2014) The gene expression database for mouse development (GXD): Putting developmental expression information at your fingertips. *Dev. Dyn.*, **243**, 1176–1186.
8. Smith,C.M., Finger,J.H., Hayamizu,T.F., McCright,I.J., Xu,J., Berghout,J., Campbell,J., Corbani,L.E., Forthofer,K.L., Frost,P.J. *et al.*, (2014) The mouse Gene Expression Database (GXD): 2014 update. *Nucleic Acids Res.*, **42**, D818–D824.
9. Davies,J.A., Little,M.H., Aronow,B., Armstrong,J., Brennan,J., Lloyd-MacGilp,S., Armit,C., Harding,S., Piu,X., Roochun,Y. *et al.*, (2012) Access and use of the GUDMAP database of genitourinary development. *Methods Mol. Biol.*, **886**, 185–201.
10. Brown,S.D. and Moore,M.W. (2012) The international mouse phenotyping consortium: past and future perspectives on mouse phenotyping. *Mamm. Genome*, **23**, 632–640.
11. Hayamizu,T.F., Baldock,R.A. and Ringwald,M. (2015) Mouse anatomy ontologies: enhancements and tools for exploring and integrating biomedical data. *Mamm. Genome*, **26**, 422–430.
12. Hayamizu,T.F., Wicks,M.N., Davidson,D.R., Burger,A., Ringwald,M. and Baldock,R.A. (2013) EMAP/EMAPA ontology of mouse developmental anatomy: 2013 update. *J. Biomed. Semantics*, **4**, 15.
13. Sunkin,S.M., Ng,L., Lau,C., Dolbeare,T., Gilbert,T.L., Thompson,C.L., Hawrylycz,M. and Dang,C. (2013) Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res.*, **41**, D996–D1008.
14. Schmidt,E.F., Kus,L., Gong,S. and Heintz,N. (2013) BAC transgenic mice and the GENSAT database of engineered mouse strains. *Cold Spring Harb. Protoc.*, **2013**, 3.
15. NCBI Resource Coordinators. (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **44**, D7–D19.
16. Petryszak,R., Keays,M., Tang,Y.A., Fonseca,N.A., Barrera,E., Burdett,T., Füllgrabe,A., Fuentes,A.M., Jupp,S., Koskinen,S. *et al.* (2016) Expression Atlas update–an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.*, **44**, D746–D752.
17. Howe,D.G., Bradford,Y.M., Conlin,T., Eagle,A.E., Fashena,D., Frazer,K., Knight,J., Mani,P., Martin,R., Moxon,S.A. *et al.* (2013) ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. *Nucleic Acids Res.*, **41**, D854–D860.
18. James-Zorn,C., Ponferrada,V.G., Burns,K.A., Fortriede,J.D., Lotay,V.S., Liu,Y., Karpinka,B.J., Karimi,K., Zorn,A.M. and Vize,P.D. (2015) Xenbase: Core features, data acquisition, and data processing. *Genesis*, **53**, 486–497.
19. Antin,P.B., Yatskievych,T.A., Davey,S. and Darnell,D.K. (2014) GEISHA: an evolving gene expression resource for the chicken embryo. *Nucleic Acids Res.*, **42**, D933–D937.
20. Motenko,H., Neuhauser,S.B., O'Keefe,M. and Richardson,J.E. (2015) MouseMine: a new data warehouse for MGI. *Mamm. Genome*, **26**, 325–330.