D860–D865 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 12 December 2007*
doi:10.1093/nar/gkm938

EMAGE—Edinburgh Mouse Atlas of Gene Expression: 2008 update

Shanmugasundaram Venkataraman, Peter Stevenson, Yiya Yang, Lorna Richardson, Nicholas Burton, Thomas P. Perry, Paul Smith, Richard A. Baldock, Duncan R. Davidson and Jeffrey H. Christiansen*

MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK

Received September 14, 2007; Revised October 10, 2007; Accepted October 11, 2007

ABSTRACT

EMAGE (http://genex.hgu.mrc.ac.uk/Emage/database) is a database of *in situ* gene expression patterns in the developing mouse embryo. Domains of expression from raw data images are spatially integrated into a set of standard 3D virtual mouse embryos at different stages of development, allowing data interrogation by spatial methods. Sites of expression are also described using an anatomy ontology and data can be queried using text-based methods. Here we describe recent enhancements to EMAGE which include advances in spatial search methods including: a refined local spatial similarity search algorithm, a method to allow global spatial comparison of patterns in EMAGE and subsequent hierarchical-clustering, and spatial searches across multiple stages of development. In addition, we have extended data access by the introduction of web services and new HTML-based search interfaces, which allow access to data that has not yet been spatially annotated. We have also started incorporating full 3D images of gene expression that have been generated using optical projection tomography (OPT).

INTRODUCTION

Techniques for assessing sites of gene expression *in situ* within an intact specimen, such as *in situ* hybridization (ISH), immunohistochemistry (IHC) or *in situ* transgenic reporter (ISR) approaches yield image-based information about spatially complex patterns of gene expression within tissues and organisms. For archiving and data retrieval, such images have traditionally been visually assessed for sites of expression by a human annotator, who subsequently manually annotates an anatomy ontology to describe the parts of the organism/tissue where expression is/is not detected. This method, whilst being excellent for warehousing the data at a relatively gross level, cannot easily describe the spatial intricacies of complex gene expression patterns. In addition, this approach can be constrained by the availability of anatomical expertise and time with which to perform these manual annotations.

For ISH, IHC and ISR studies in the developing mouse embryo, we have developed a parallel approach for archiving *in situ* expression data whereby spatial information regarding the sites of gene expression as captured in data images is transferred directly into a spatial atlas of mouse development (EMAP) and then housed in an accompanying database (EMAGE) (1–3). Advantages of this spatial annotation approach include that instances of data with incomplete or absent textual annotations can be retrieved from database searching, and users with little or no prior knowledge of anatomy also gain access to the data. Within EMAGE, sites of gene expression are also described using a mouse embryo anatomy ontology, thus it is possible to search EMAGE using a flexible combination of spatial- and text- based methods. The standardized spatial- or text- annotations housed in EMAGE are always supported by digital images of the raw data as well as detailed information regarding the probe/antibody/transgenic line construction, the specimen (strain, age, etc.) and experimental methods (e.g. fixation, visualization technique).

EMAGE is implemented in an Object Store object-oriented database management system. Full-time editorial staff adds new spatially annotated data regularly. Database access is through a combination of HTML-based and Java Client Interfaces which have been described in detail elsewhere (3). This article outlines recent increases in data coverage, advances in spatial searching methods and results visualization.

DATA CONTENT

Spatially annotated data

Each spatially annotated EMAGE entry contains at least one original data image along with the accompanying

*To whom correspondence should be addressed. Tel: +44 131 332 2471; Fax: +44 131 467 8456; Email: jeff.christiansen@hgu.mrc.ac.uk

spatial mapping. Section data is mapped into the 3D space of a reference model and whole-mount (WM) data is mapped onto a 2D projection image of a 3D reference model. At least one 3D reference model exists for all stages between TS07-20. More than 95% of spatially annotated entries also include an accompanying text-based annotation. As of September 2007 there were 3656 entries with spatial annotation covering 1483 genes. Of these, 2935 entries correspond to data from wholemount and 721 from sectioned samples. See Supplementary Data (Table S1 and Figure S1) for a further breakdown on the basis of stage of development. Of all entries, 3663 entries correspond to ISH data, 109 to IHC data and 4 to ISR data. 8% have been direct submissions to EMAGE from contributing labs, 52% is data that has been previously published in the literature and 40% from screening consortia.

A further 316 entries have been submitted directly to EMAGE from individual labs and spatially annotated. These will be made publicly available upon publication of the results in the literature.

EMAGE curators have recently begun to score the quality of incoming data images and the degree of morphological similarity between the data specimen and target template for spatial mapping. This is achieved using a simple three-level ranking system (good, moderate, poor: see http://genex.hgu.mrc.ac.uk/Emage/database/EMAGE_Docs/Key_to_spatial_rankings.html for further information). These scores have also been retrospectively assigned to all previous spatial annotations in the database. They can be used to gauge the potential quality of each spatial annotation and for filtering data sets for spatial analyses as discussed below (such that only the highest quality annotations are used, for example).

Non-spatially annotated data

In addition to the spatially annotated entries discussed above, users of EMAGE have access to an extensive set of further data images. These are currently obtained from the literature, as well as two new sources: the EURExpressII consortium (http://www.eurexpress.org) and a NIH-funded craniofacial screening consortium (NIDCR P50 DE016215-01).

Literature

Data from the literature continues to be incorporated with the help of our colleagues at the GXD database (4) who identify and then text annotate the data using the EMAP anatomy ontology (5). EMAGE staff assess all of the images held in GXD (for their spatial mapping suitability) from several journals for which we have been granted copyright permissions from the Publishers to reproduce the images (Mechanisms of Development, Gene Expression Patterns, Developmental Biology and Development).

EURExpressII

EURExpressII is a European Union funded consortium currently producing sectioned ISH data for all mouse genes on 14.5 days *post coitum* (d*pc*) embryos (25 sagittal sections per gene). As this data is generated, it is made publicly available via EMAGE.

NIH (NIDCR) craniofacial data

Data from the NIDCR (National Institute of Dental and Craniofacial Research) funded craniofacial screening consortia is produced as WM ISH data at four stages of development (9.5, 10.5, 11.5 and 12.5 d*pc*). In addition to digital photography, most data embryos from this source are being imaged at the MRC Human Genetics Unit (Edinburgh, UK) using Optical Projection Tomography (OPT), a technique that yields a 3D image (6). EMAGE has been extended to hold raw OPT data images [in woolz 3D digital image format (7)], along with associated movie visualisations (of the rotating specimen, plus all sections along the *XY*-, *YZ*- and *ZX*- section planes from the reconstructed 3D object) which allow easy data browsing of the 3D objects. The 3D woolz format images can be downloaded and interactively viewed using our bespoke software, JAtlasViewer [(8) http://genex.hgu.mrc.ac.uk/Software/JavaTools/JAtlasViewer/intro.html]. For an example EMAGE entry containing OPT data, see http://genex.hgu.mrc.ac.uk/das/jsp/submission.jsp?id=EMAGE:3837. See on-line Supporting MovieS1 for a screencast movie showing how to browse the content of an OPT entry.

Access to non-spatially mapped data

In addition to the data spatially annotated by our curators, non-spatially mapped data is accessible via an EMAGE 'repository' (accessible by browsing from the EMAGE homepage or directly from http://genex.hgu.mrc.ac.uk/Emage/rep/index.html, see Figure 1). All data in the repository is indexed by a number of criteria and can be searched by: *Source* (e.g. *EMAGE, EURExpressII*, etc); *ID* (formatted as EMAGE: for spatially annotated data and EMAGE:R for non-spatially annotated data); *gene*/*protein symbol*; *Reagent ID* of the probe/antibody used; *stage of development* (both Theiler stage and other staging systems); *genotype* and *assay type* (ISH/IHC/ISR). Thumbnail images of the original data are also shown in this table. Columns can be ordered in ascending or descending alpha-numerical order and simple text searches, including the use of wildcards can be used to search the repository. Gene name/symbol synonym searching is also supported. Any number of entries can be selected and a 'collection' made that can be retrieved at a later time through the use of cookies. EMAGE:R IDs are directly linked to the data in the original data source. See online Supplementary Movie S2 for a screen-cast showing how to browse the 'repository' data tables.

The repository gives EMAGE users access to a further 24 610 assays for 8597 genes (~250 000 images). Thus, in total around ~30 000 assays (with ~250 000 images) for ~10 000 genes are available in a unified format via EMAGE.

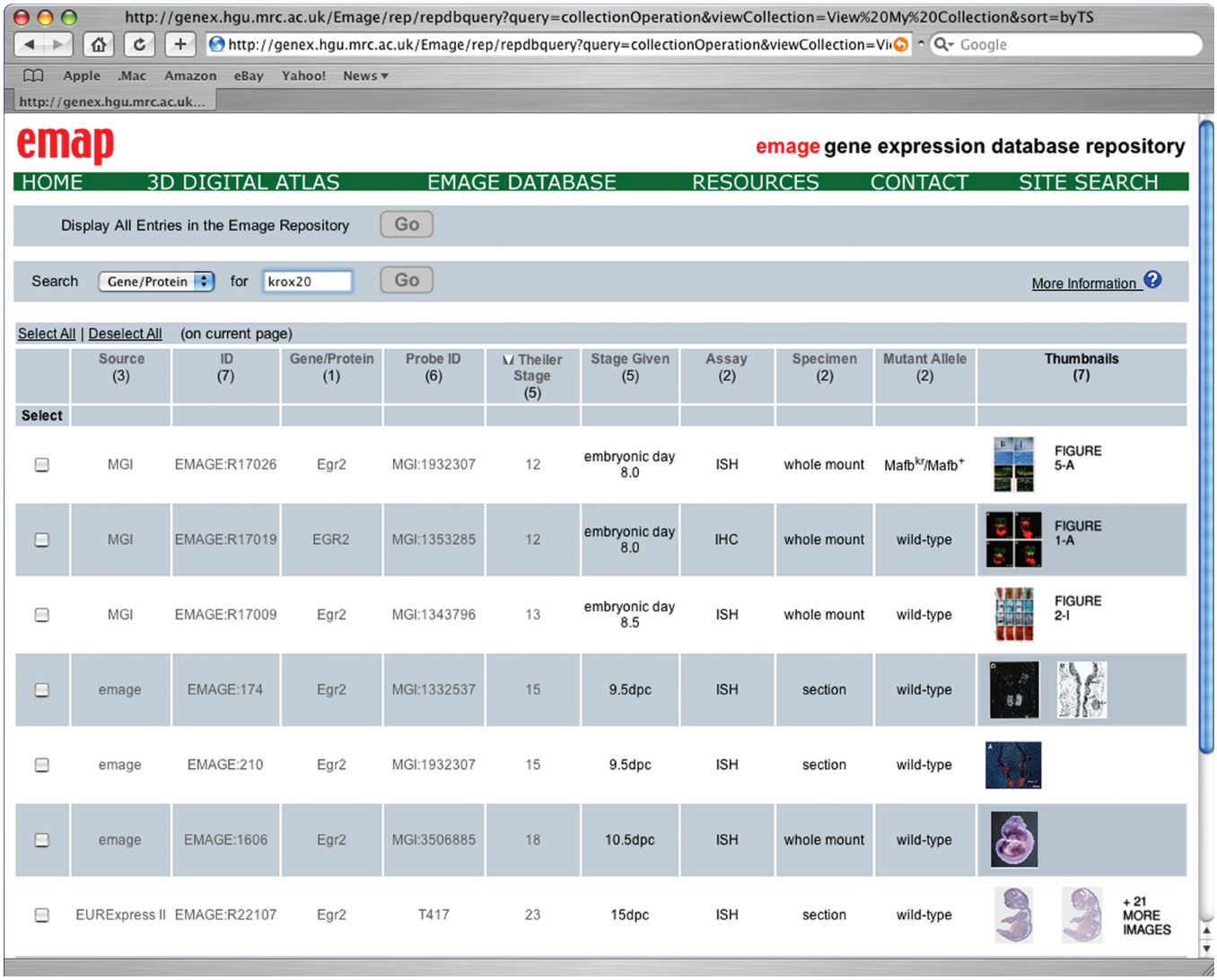

Figure 1. Example results table from the EMAGE Data Repository. The results from a Gene/Protein search of the EMAGE repository for the unofficial gene symbol 'krox20' is shown. Results are retrieved and listed under the currently approved symbol for Krox20: Egr2. The list has been ordered by ascending Theiler Stage value. Examples of data from a number of original sources, for ISH and IHC as well as wild-type and heterozygous mutant specimens are shown. Thumbnail images are linked to the full-size originals.

ENHANCED SPATIAL DATA ANALYSIS CAPABILITIES

One of the main foci in the recent development of EMAGE has been to extend searching capabilities based on the spatial annotations. These fall broadly into three areas: searching across multiple stages of development, data mining via global pattern similarity searching and spatial similarity searching over a localized region of the embryo.

Multiple stage searching

Until recently, spatial searches in EMAGE have been restricted to one TS model. This has now been developed to allow spatial searches across multiple stages of development within a subset of the available EMAP reference models (i.e. WM models TS15–TS19). To achieve this, between 60 and 90 points have been defined at roughly anatomically equivalent points between temporally adjacent reference models (see online Supplementary Figure S2 for an illustration of these points). The spatial transformations that are defined by these points between the models can then be applied to data in the database to allow spatial searching of data that has been mapped onto several reference models. In a similar manner, transformations between left and right WM views of every EMAP model are also now possible. These spatial transformations have been used in both the global pattern similarity searching and the spatial similarity searching over a localized region of the embryo discussed below.

Global pattern similarity searching

We have recently added functionality such that large image sets of WM and 3D data at each stage of

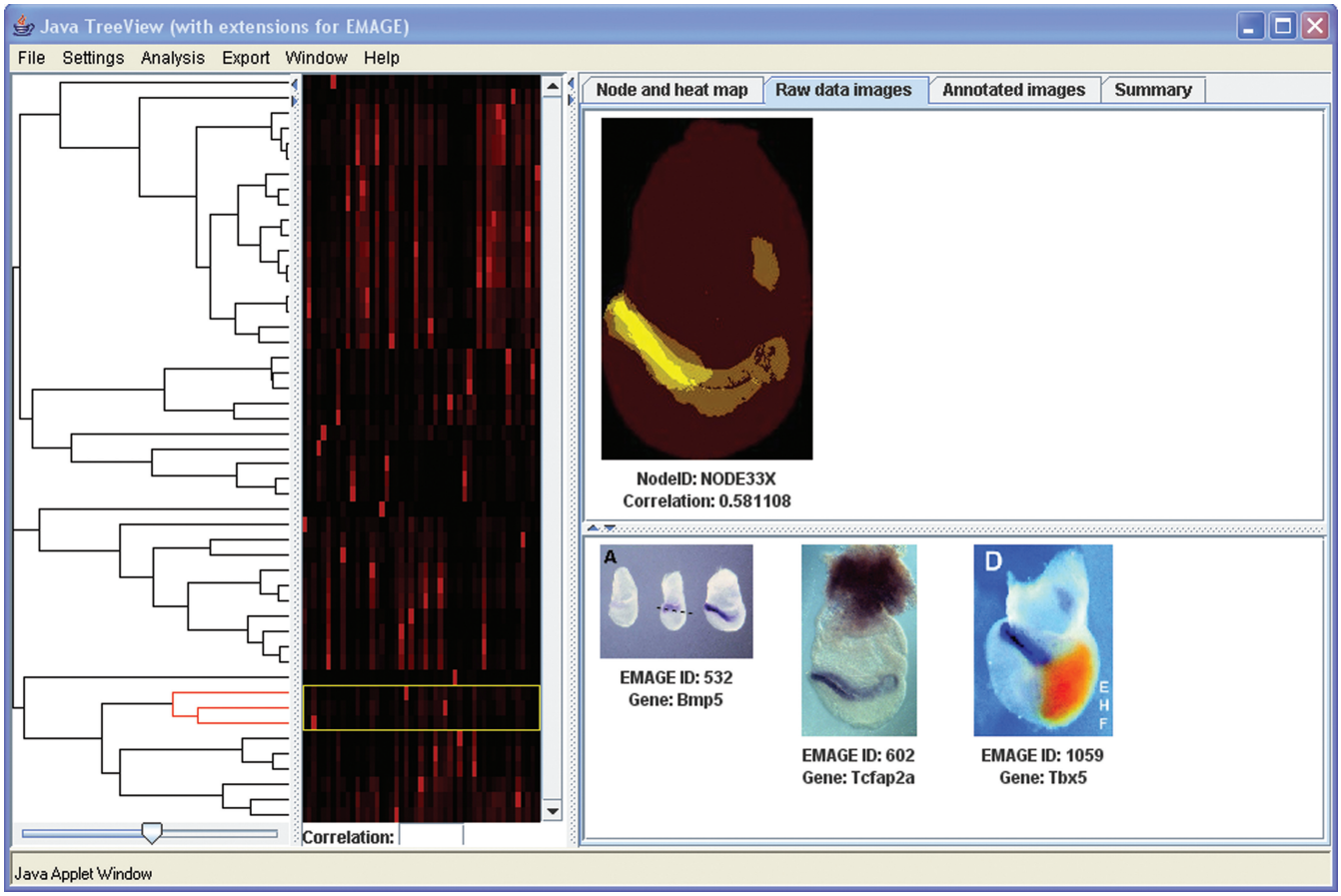

Figure 2. Display of spatial pattern comparison hierarchical clustering results in the JavaTreeView Applet. An example dataset of 49 images of WM TS11 embryos showing a wide variety of different expression patterns was selected for spatial comparison and subsequent hierarchical clustering analysis. The results are displayed in a version of JavaTreeView modified to run as an applet and with the ability to load images. The dendrogram and associated matrix (which offer the user information about pattern relatedness) appear on the left. On the right, visual feedback of the common spatial pattern found in all annotations on the selected branch is shown in the form of a 'heat-map' (yellow on black). In this case, the 'raw data images' tab is selected to show the 3 images that have contributed to the pattern in the heat map (expression in the images is in blue/purple).

development (or over several stages of development) can be selected by a user and then compared to each other based on their spatial similarity. These tools allow interactive exploration of the data and can be used to find examples of clusters of images that display globally similar expression patterns, which may point to the existence of common regulation or synexpression groups (9). These would otherwise have to be found by eye, which is becoming increasing unfeasible as available datasets become larger. This functionality can be accessed by browsing from the EMAGE homepage, or directly from http://genex.hgu.mrc.ac.uk/Emage/cluster_analysis/index.html

The method requires a user to select a dataset of interest. All images in that dataset are then displayed. Subsequently the user is prompted to select the regions denoting apparent signal intensity levels (strongest, moderate, weakest) to include in the calculations [see (3) for further information on signal intensity levels]. Spatial similarity comparisons have been pre-calculated between every possible pair using the Jaccard Index (V), which is defined as $(d_1 \cap d_2)/(d_1 \cap d_2)$, where d_1 and d_2 are the two input spatial domains. Hierarchical clustering of the pair-wise comparison results has also been pre-calculated using Cluster 3.0 software (10), using un-centred correlation similarity metric followed by complete linkage clustering and cluster output files are generated (in. cdt format).

Users have several options to visualize the clustered data presented to them. The most basic is to view the original data images in order from left to right as they appear in the output cluster files from top to bottom. This method groups images together that display similarity but gives no visual feedback of the hierarchical tree. Alternatively, a user may select to launch an interactive TreeView applet (or an application version of the same software) that has been developed from the open-source software Java TreeView 1.0.13 (11). The EMAGE version allows a user to peruse the tree and view 'heat map' representations of sites of expression in the relevant embryo reference model that are contributed from images on each branch. In addition, the contributing raw data images and spatial mapped representations for the selected branch may be viewed. A slider tool is available to select multiple branches from the tree, and the tree can also be searched for examples of a gene of interest or EMAGE:IDs. See Figure 2 for a screenshot of the

JavaTreeView Applet interface and online Supplementary Movie S3 for a screencast depicting how to perform a search of this type.

Spatial similarity searching over a localized region

Similarity searching has also been extended to allow user-defined, spatial searches over a localized region of the embryo (*vis-à-vis* global patterns as described above). This employs a new algorithm called LOSSST (*L*ocal *S*patial *S*imilarity *S*earch *T*ool), which is conceptually similar to the BLAST (*B*asic *L*ocal *A*lignment *S*imilarity *T*ool) algorithm (12) used in local comparisons of nucleic acid sequences. The method requires the user to first define an arbitrary search region in the Java Client Interface. LOSSST then defines a local region for spatial comparison by dilating 30 pixels/voxels in all directions from the edges of the user-defined original query domain. Jaccard similarity scores are then calculated between the query domain and all domains in the database within the local comparison region. The search results are ordered from best spatial fit to least and allows a user to retrieve images that display spatial similarities in gene expression patterns over localized regions of the embryo. This search can also be performed over multiple stages of development (see online Supplementary Movie S4 for a screencast capture movie depicting this type of search being perfomed in the Java Client).

DATA ACCESS

Data access has been improved over the past several years by the introduction of EMAGE web services which allows direct access to the EMAGE database server over the internet, using a software client. The interface to the web services is described by WSDL (http://www.w3.org/TR/wsdl), which defines the services provided and the data structures involved. Programmers can use the WSDL description to design client software that makes requests to EMAGE and processes the results that are returned. The EMAGE database uses Apache Axis (http://ws.apache.org/axis) to deliver its web services.

FUTURE DIRECTIONS

EMAGE will continue to source and spatially annotate data in the developing mouse embryo. This includes data from the literature, which we continue to annotate in conjunction with our colleagues at the GXD, as well as data from screening consortia.

As most of our data has thus far originated in many different labs, and has been imaged in an *ad hoc* manner, the current spatial annotation procedure requires significant human input to assess both the Theiler stage of development and the specimen view depicted in each image, and to then manually define points of equivalence between the data and reference images. As we gain access to large, consistently produced image datasets such as EURExpressII, this will allow the use of computational methods for spatial integration of these data and significantly reduce the amount of 'hands-on' time required. These are currently being explored.

We are also actively investigating methods to perform 3D transformations of one volume into another. This will allow 3D data imaged using OPT to be spatially incorporated into the EMAGE framework, and spatial transformations to be defined between pairs of 3D reference models, which will allow spatial-based searching in 3D across multiple stages of development, similar to the 2D WM searching across multiple stages as discussed previously.

Search capabilities that are currently only available via the Java Interface Client (e.g. spatial searching) will be moved to HTML webpage interfaces in the near future, to allow more user-friendly searching of EMAGE. The Java Interface Client will be retained as a program that can be used to create a local private database for in-lab data management, and to electronically submit data to EMAGE for curation and inclusion in the public database.

Finally, we are currently in the process of re-factoring EMAGE from an Object-oriented database model (using ObjectStore) to a relational database model (using DB2). This change will be completed in the near future.

USER HELP

There is dedicated User Support for EMAGE. Please write to ma-edit@hgu.mrc.ac.uk

We also have a User Group where we announce new releases and other relevant information. To subscribe, visit http://www.jiscmail.ac.uk/lists/MA-EMAGE.html

CITING EMAGE

To reference EMAGE, please cite this article. For specific data entries, please list the EMAGE:ID and also mention that the data was retrieved from EMAGE, MRC Human Genetics Unit, Edinburgh, UK (http://genex.hgu.mrc.ac.uk).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

EMAGE is supported by the Medical Research Council. We would like to thank Dr Martin Ringwald and our colleagues at the GXD (MGI, Jackson Laboratory, Maine, USA) for input; The Company of Biologists Ltd. and Elsevier B.V. for copyright agreements allowing reproduction of images from Development, Developmental Biology, Mechanisms of Development and Gene Expression Patterns and Mehran Sharghi for technical help in development of the TreeView Applet. Funding to pay the Open Access publication charges for this article was provided by the Medical Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Davidson,D. and Baldock,R. (2001) Bioinformatics beyond sequence: mapping gene function in the embryo. *Nat. Rev. Genet.*, **2**, 409–417.
2. Baldock,R.A., Bard,J.B., Burger,A., Burton,N., Christiansen,J., Feng,G., Hill,B., Houghton,D., Kaufman,M. *et al.* (2003) EMAP and EMAGE: a framework for understanding spatially organized data. *Neuroinformatics*, **1**, 309–325.
3. Christiansen,J.H., Yang,Y., Venkataraman,S., Richardson,L., Stevenson,P., Burton,N., Baldock,R.A. and Davidson,D.R. (2006) EMAGE: a spatial database of gene expression patterns during mouse embryo development. *Nucleic Acids Res.*, **34**, D637–D641.
4. Smith,C.M., Finger,J.H., Hayamizu,T.F., McCright,I.J., Eppig,J.T., Kadin,J.A., Richardson,J.E. and Ringwald,M. (2006) The mouse Gene Expression Database (GXD): 2007 update. *Nucleic Acids Res*, **35**, D618–D623.
5. 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.
6. Sharpe,J., Ahlgren,U., Perry,P., Hill,B., Ross,A., Hecksher-Sorensen,J., Baldock,R. and Davidson,D. (2002) Optical projection tomography as a tool for 3D microscopy and gene expression studies. *Science*, **296**, 541–545.
7. Rutovitz,D. (1968) In Cheng,G.C., Ledley,R.S., Pollock,D. K. and Rosenfeld,A. (eds), *Pictorial Pattern Recognition*, Thompson Book, WA, pp. 105–133.
8. Feng,G., Burton,N., Hill,B., Davidson,D., Kerwin,J., Scott,M., Lindsay,S. and Baldock,R. (2005) JAtlasView: a Java atlas-viewer for browsing biomedical 3D images and atlases. *BMC Bioinformatics*, **6**, 47.
9. Niehrs,C. and Pollet,N. (1999) Synexpression groups in eukaryotes. *Nature*, **402**, 483–487.
10. Eisen,M.B., Spellman,P.T., Brown,P.O., and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
11. Saldanha,A.J. (2004) Java Treeview–extensible visualization of microarray data. *Bioinformatics*, **20**, 3246–3248.
12. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*, **215**, 403–410.