

Report of EMAGE advisory group

MRC Human Genetics Unit
Edinburgh, UK
30th November, 2007

Advisory Board members present:

Dr. David Wilkinson – National Institute for Medical Research, UK (Chairperson)
Dr. Janan Eppig - Mouse Genome Informatics, USA
Dr. Graham Kemp – Chalmers University of Technology, Sweden
Dr Suzanna Lewis – Lawrence Berkeley National Laboratory, USA
Dr. Helen Parkinson – European Bioinformatics Institute Hinxton, UK
Dr. Martin Ringwald – GXD Database, Mouse Genome Informatics, USA
Prof. Claudio Stern – Department of Anatomy, University College London, UK
Dr. Sarah Wedden – MRC Technology, UK

Apologies:

Prof. Steve Brown – MRC Mammalian Genetics Unit, Harwell UK.

The aim of the EMAGE project is to establish a database of gene expression spatial patterns mapped onto an annotated three-dimensional atlas of mouse embryo anatomy. The EMAGE database is being developed in close collaboration with GXD (Jackson Laboratory, USA) which provides a complementary text-based database of gene expression patterns, together with other information. EMAGE and GXD are being further coordinated as components of the Mouse Gene Expression Information Resource (MGEIR). There are two main aspects to the development of the EMAGE database: (A) Entry of gene expression patterns into the database; (B) Software development, including interfaces, display and search tools that facilitate data mining. EMAGE is being proactively publicised and coordinated with databases being established for other organisms.

RECOMMENDATIONS

(A) Data Entry

The representation of a large number of genes is a key requirement for the database. Since the advisory board meeting in Nov 2005, there have been ~1500 new entries with spatial annotations corresponding to ~520 genes (now ~3900 entries and 1513 genes in total). As recommended by the advisory board, non-annotated images of gene expression patterns have been included in the database, and these now comprise ~230,000 images corresponding to ~10,000 genes. OPT images of spatial expression are now being entered into EMAGE. In addition to the published literature, a number of screens of gene expression data have been made available to EMAGE.

While substantial progress has been made, the rate of entry of spatially mapped patterns has been considerably lower than the target of 1000-1500 per year. This is in part due to there being a low number of staff available for carrying out data entry, who necessarily also spend some time contributing to other aspects of database development. It is essential that there is a sustained increase in the rate of data entry during the next 12 months and onwards, of at least 1500 genes per year. In addition, it is important to focus efforts in order to increase the value of EMAGE more rapidly. We have the following recommendations to help achieve this:

1. Develop strategies to increase the rate of annotation:
 - a. By streamlining the process to allow for batch entry of data from the same tissues / section plane / stages.
 - b. Based on previous experience, it may also be possible, by looking for correlations in the literature annotation that has already been carried out by GXD, to carry out a more rapid triage and avoid spending time on published data that lack required information, are of poor quality, or are similar to data already present in EMAGE.
 - c. Explore the possibility of developing semi-automated methods to triage the images and speed up the annotation process.
 - d. To provide feedback to ENSEMBL when gene tracking problems are uncovered.

2. That data that add most to the value of EMAGE to the scientific community should be entered first. In general, there will be greater value to entering data for more genes at fewer stages, rather than for fewer genes at more stages; this may increase the likelihood of discovering novel genes that have spatial overlaps with known genes suggestive of functional or regulatory relationships. This can be achieved by focussing data entry on selected developmental stages, such as 9.5 dpc and 14.5 dpc. In particular, the forthcoming entry of mapped data from EURExpress II on sections of 14.5 dpc embryos will help to fulfill this objective, and these data should be entered as soon as possible.

(B) Software Development

Impressive progress is being made with developing powerful new tools such as cluster analysis of the similarity of spatial expression patterns. The plans to increase the user-friendliness of the interface are very good. We recommend that priority is given to the following aspects of software and interface development:

1. That the front page of EMAGE be made to state the specific value of this database and of GXD (what sort of information can be mined from each

- of them). Similar information should be on the front page of GXD and MGEIR.
2. It will be very helpful to provide an example of novel findings from data mining from EMAGE as soon as such results are available, preferably in the next 6 months.
 3. The website polishing described in the EMAGE report should be a high priority.
 4. Identify the user community for EMAGE and GXD by survey etc to improve support of the needs of specific user communities
 5. To provide high profile user documentation/canned queries in any new GUI for e.g. clustering for genes of interest as a demonstration of EMAGE capabilities distinct from GXD
 6. Different probes to the same gene that show expression in non-overlapping areas should be queryable.
 7. Measure how frequently users are unable to find what they are looking for and take corrective steps if possible.
 8. Add tutorial examples for the new interfaces to the web site. Each example should include an explanation of why the analysis might be done, in addition to how it's done. This should also help non-specialists outside the mouse development and gene expression communities to appreciate the usefulness of EMAGE.
 9. Add glossary information (for "Theiler Stage", "ISH", "dpc", etc.) that can be accessed conveniently from the web interface.
 10. Information on how to cite EMAGE should be more prominent on the web site.

(C) Publicising the database

Establishment of the EMAGE database as an internationally recognised resource is continuing to be promoted by presentations at a number of meetings and courses.

1. These efforts should continue, for example running courses at EBI and participation in the Cold Spring Harbor Mouse course. This should be a priority once new user interfaces are developed.
2. As soon as the relevant work has been done, the EMAGE group should submit a peer-reviewed paper for publication with a summary of what can be done by using the database and search facilities, and include at least one example of something interesting discovered from this alone. This publication could be a review or a primary paper or a hybrid.
3. To increase awareness of the value of EMAGE and GXD by a wider community, efforts should be continued to establish two-way links with other databases such as OMIM. Efforts should also be continued to ensure that following updates of other databases they maintain their existing links with EMAGE.

(D) Relationship with other gene expression databases

In the longer term, coordination and links between gene expression databases of different species is essential for comparative analysis, for example to find conserved and divergent expression patterns. Interactions are being maintained with a human gene expression database project, and EMAGE members are involved in a grant proposal for a chick gene expression database. The interactions with these other systems should continue.

(E) Scientific Advisory Board meetings

To ensure that progress is monitored and timely feedback provided, the next Advisory Board meeting should be in one year's time.

ACTION POINTS

1. Develop appropriate strategies and achieve >1500 new entries per annum (recommendation A1) (12 months).
2. Prioritise the entry of the most useful data (A2) (start immediately).
3. Place additional basic information on website (B1, B9, B10) (1 month).
4. Establish example of use of database, place on website and and prepare manuscript for submission (B2, C2) (9 months).
5. Polishing of website as described in report (B3) (3 months).
6. Addition of further information and capabilities to website, including from results of user surveys (B4, B5, B6, B7, B8) (6 months)
7. Continue efforts to publicise database (C1, C3)
8. Continue interactions with databases on other organisms (D)
9. Organise next advisory board meeting in one year.