

Minutes from the 2005 EMAGE Advisory Board Meeting
18th November 2005
MRC Human Genetics Unit
Edinburgh

Advisory Board members present:

Dr. David Wilkinson – National Institute for Medical Research, UK. (Chairperson)

Dr. Graham Kemp – Chalmers University of Technology, Sweden

Prof. Claudio Stern – Department of Anatomy, University College London, UK

Dr. Sarah Wedden – MRC Technology, UK

Dr. Alvis Brazma – European Bioinformatics Institute Hinxton, UK

Apologies:

Dr. Martin Ringwald – GXD Database, Mouse Genome Informatics, USA

Dr. Janan Eppig – Mouse Genome Informatics, USA

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

EMAGE members present:

Prof. Richard Baldock – Principal Investigator, EMAP project, HGU, UK.

Dr. Duncan Davidson – Principal Investigator, EMAP project, HGU, UK.

Dr. Jeff Christiansen – Project Leader, EMAGE Database, HGU, UK.

Agenda:

10:00 Welcome

10:15 EMAGE progress report for 2004/5 (presentation by Jeff Christiansen)

- *Brief reminder of the concepts of the EMAP Atlas and EMAGE gene expression database and the aims of the EMAGE project.*
- *Outline of staff specifically in EMAGE as well as those in the entire EMAP project and how available staff are deployed with respect to EMAGE*
- *Recap of state of EMAGE at 2004 Advisory Meeting.*
- *Advances made since the last Advisory Meeting – e.g. data entry rate, interface and functionality changes, securing of large datasets for future entry, community outreach.*

1:00 Lunch

2:00 Immediate and long-term aims for EMAGE (presentation by Jeff Christiansen)

- *Outline of current changes underway for EMAGE*
- *Discussion of the long-term aim of EMAGE:*
 - *Who is/are our target audience/s?*
 - *How much effort should be devoted to extending query functionality vs. entering large data volumes?*
 - *How to increase usage among out target audience/s?*

4:30 Drafting Report and Action Points for 2006.

5:00 End

Notes from the discussion in response to Jeff Christiansen's presentation on the EMAGE progress report for 2004–2005 and Potential changes for 2005–2006:

EMAP ATLAS

- Extend atlas to include later stages
- Mapping (spatial and anatomical) between stages is important
- Painting is important – educational value and also to integrate spatial and textual annotations.

EMAGE

- html access to database is important – develop this for all types of query, except where not possible (e.g. Q on arbitrary regions).
- Claudio Stern – how do you annotate expression in migrating N.C. cells? – answer: anat. ontol. should have neural crest–derived cells as a component, and a parent of this should be painted; therefore painted anat domain contains n.c. cells; therefore expression in, or intersecting with, this painted anat domain may be present in n.c. cells. The problem is then the resolution of anatomy detailing where n.c. cells are present and the resolution of painting.
- Claudio Stern – the system should guide the inexperienced user to ask the right questions of it in order to help them map their data. At points in the work flow, what questions should the user be asking and how can the system help to provide the answers. Same for query. If this can be done it will help to educate the inexperienced user. User Kit for the kit-oriented molecular biologist!
- Alvis Brazma – people want tools that make sense of the list genes returned from a query. What can the DB tell me about my gene? Links to GO?
- Claudio Stern – what do these genes have in common?
- Graham Kemp – important to be able to annotate at the cell-type level. All agreed that cells should not be resolved in the anat. ontology. Implement links to a cell-type ontology for additional annotation. Eventually, could add links from anat. leaves to cell-type ontology.
- Claudio Stern – Cross-stage queries are important.
- Claudio Stern – integration of text and spatial queries is crucial
- Claudio Stern – WM data: left + right queries should have option L or R.
- New webpages should say 'demo' rather than 'preview'. Html query pages should have comment 'this is not a spatial search.' Html page header pointing to main EMAGE page should be more prominent.
- The Advisory Board report should go to someone in Head Office and should be sent to Nick Hastie.
- Claudio Stern – Chick database suggested Roslin to host DB, or UCL, Kings or Dundee. Later discussion led Claudio to propose that the DB would be best held at the HGU for the sake of efficiency and updating with current developments in EMAGE.
- Duncan Davidson asked how should EMAGE deal with data that cannot, for whatever reason, be spatially mapped? Discussion resulted in agreement that this data should be held in a repository of uncurated, unmapped data alongside EMAGE and developed at the HGU. Considering this compartment of unmapped data, there would be different types – data that's mappable but not yet mapped; textually-annotated, unmappable data and unmappable, unannotated data that is nevertheless good quality. Need to collaborate where there is overlap here with the GXD, but with GXD data is that there is often no direct access to individual images (only composites).
- David Wilkinson – Look at ZFIN as a good example of links between gene name and image of expression pattern.

- Alvis Brazma mentioned SwissProt and TrEMBL as example of how we might deal with curated versus uncurated data.
 - Alvis Brazma suggested exploring a portal approach to query >1 DB
 - Discussed quality/confidence indices. Should we include 'screen vs non-screen' annotation? Alvis Brazma suggested colour code – don't give too many grades of confidence or too many choices. Look at confidence indicators in the GO annotations.
 - Show data with different levels of confidence in different columns or (Richard Baldock) just rank by confidence.
 - L/R images of WM – we need to check with authors (e.g. McMahon) that the image is not a mirror image. This level of curation is important in EMAGE.
 - Follow up ex-students of EMAGE courses for comments and are they using it?
 - Alvis Brazma – google-type search. Gene name; tissue name. Discussed whether one field or up to 3.
 - Use simple and advanced search modes
 - Sarah Wedden – EMAGE/EMAP can use Bioptronics display software (available for Linux).
 - Alvis Brazma – EBI tree viewer can be used in html to display clusters.
 - Probe database – should we map to ENSEMBL or put in NCBI. Advice was probably not yet.
 - Courses – David Wilkinson – apply to CSHL again
 - Claudio Stern – and Woods Hole (suggested contact Richard Harland)
 - EMBL courses – find out what is planned
 - Alvis Brazma – mentioned an EMBL effort to integrate gene expression data from different species (3D, gene x tissue type x species)
 - Richard Baldock – the XSPAN Project aims to integrate anatomy ontologies for different species.
 - Claudio Stern – ZF – contact Fons Verbeek to find out how ZF atlas is progressing.
 - Claudio Stern – Key question is: who is the target audience for EMAGE?
 - Jeff Christiansen – mouse developmental biologists who want to use spatially mapped data.
 - Claudio Stern – suggests 3 kinds of user:
 1. Molecular Biologists who are not expert in developmental biology
 2. Medics who want to look for candidate genes or do virtual experiments with the data
 3. expert developmental biologists

1 and 2 are novices as far as the mouse embryo is concerned. 3 are experts. The functionality needed by the novices and experts is very different.

Claudio Stern and Alvis Brazma – The novices need simple interfaces (like Google)

Claudio Stern – but sophisticated novices may want to test sensible hypotheses using quite advanced methods but not necessarily knowledge of developmental biology.

Alvis Brazma – yes, but make it simple first.

David Wilkinson – we need a balance between the two – especially, we need Boolean queries and tree/clustering) analyses.
 - Strategies to stem boredom in curators: Alvis Brazma – in EBI we encourage curators to devise ways to upload batch submissions – e.g. using perl scripting.
- Back to discussion about curated data versus repository:

Alvis Brazma – Array data at EBI is held in (A) repository (dump) and (B) warehouse (curated data). (A) takes online submissions – small submissions directly from users, large submissions with help from curators as above – also pipelines from collaborators.

Graham Kemp – to help get submissions, make users submit their data before they can analyse it.

Alvis Brazma – that doesn't work for us – for array data – users can often do powerful clustering on their own machine.

- Back to queries and what users want: Alvis Brazma – users might want to take the expression of a gene and ask – what are the 5 most similar expression patterns?
- Duncan Davidson asked how many genes will we need to have in the database in, say, 5 years in order to be useful? Say 10K but depends on what kind of analysis.
- Important uses will be
 1. to find markers (for tissues or cells)
 2. to answer – what other genes are in a given syn-expression group
 3. compare expression across species as a means to explore genetic control of expression – Claudio Stern – control elements are going to be a major aspect of analysis.
- Claudio Stern and David Wilkinson – don't restrict the database to a few particular stages – temporal patterns will be important. Also, one has to go where the data is – make use of data from large-scale screens.
- Human versus automated curation: Alvis Brazma – what's the possibility of automatic curation and annotation of images?
Richard Baldock – not at present. Natural variation is an issue – maybe in 10years?
Alvis Brazma – still, what you are doing now will be valuable – it will provide the infrastructure for any future automatic curation.

- Getting recognition for EMAGE

Graham Kemp – how do you turn users into citers?

EMAGE needs to help users to make citations. We should have a clear guide to how to cite EMAP and EMAGE and every page on our website should have a reasonably prominent link to it.

David Wilkinson – one measure of success would be a demonstration of the use of our more advanced analysis and query capabilities. Also a demonstration that people are actually using spatial mapping.

Richard Baldock – advanced analyses (clustering ?) in html ?

- What do we need?
- David Wilkinson –
 - Boolean
 - Clustering
 - Google-type search
 - Ranked returns from the query 'which expression patterns are most similar to a given input pattern'
 - Repository of unmapped and uncurated expression patterns.
- Claudio Stern – the danger with the latter is that it will become a dustbin. What you need in this repository is data that is partly curated in order to have quality control, but can't be mapped.
- David Wilkinson – it's still useful to have the data somewhere.
- Jeff Christiansen – we should build a repository and work through it marking quality controlled data – probe data can be automatically quality controlled.
- David Wilkinson and Richard Baldock – OK have selected data (e.g. trustworthy screens) in the repository.
- Jeff Christiansen – then perhaps in a year or so, review the data in the repository and then perhaps get other people to submit more data.
- Duncan Davidson – should we still take both sections and wholemount data? The latter are difficult to interpret.
- Claudio Stern and David Wilkinson – take both – take what you can get – early stages will tend to be WM, later stages sections.
- Should we have additional editors in the scientific community – no would not get a good response and would reduce consistency of curation.
- Quality control: don't need to increase quality, but don't reduce – at least not more than slightly and then only if we would get great benefit in terms of

quantity. One justification for having a highly curated database is the education of users – by relating the data to expert knowledge e.g. anatomy.

- Generally, the quality problem may be solved by having the repository database.
- Advertise the repository and make sure people know about it from key points of access.
- Other species? Claudio Stern and David Wilkinson – seek further funding, specifically for the chick.
- EMAGE Leaflet
 - Claudio Stern – language should be more direct and simple. Possibly more folders (pages) with a quick guide to use – flow chart for user then ‘to do that, click here..’
 - Include a ‘mission statement’, what EMAGE can be used for now, a strategic view.
 - Jeff to draft leaflet and send around Advisory Group for comments.
 - Interface should have a blast button so that users can blast GeneBank with probe sequence.

David Wilkinson made notes of the action points.