



**Priorities for mouse functional genomic research across Europe**

**Expert Group on Improved Resources**

**Notes of meeting**

**2 September 2005, Rome**

**June 2006**



## INDEX

Background .....	1
Business.....	1
Priorities .....	1
Present.....	1
Next steps agreed at the meeting.....	2
Database resources .....	3
Improving Search facilities .....	3
Sources of information on existing databases.....	4
Curation.....	4
Annotation.....	4
Submission of data.....	4
Standards for data.....	4
Stresses.....	5
Areas for research funding .....	5
Ontologies .....	5
Ontology working group .....	5
Combining ontologies .....	5
Communication standards.....	6
Linking databases.....	6
Biological Resources.....	6
Biological resources.....	6
Linking resources .....	7
Submission of samples/mice to accessible archives .....	8
Methods for sustainability.....	8
Funding for infrastructures.....	8
Actions agreed at the meeting.....	9
Information gathering.....	9
Suggestions for co-ordination activities.....	10
Areas for potential research funding.....	10
Glossary.....	11



---

## **Background**

PRIME will consider and suggest priorities for mouse functional genomics research in Europe over the next 10 to 15 years. These recommendations will be discussed with policy makers and funders of research in the European Commission and member states, with the aim of determining the most effective ways of supporting this research and associated infrastructures.

This will pave the way towards the instigation of the European Research Area (ERA). A key factor in supporting the research is the successful storing and retrieval of data generated in the research and the physical maintenance of resources, such as mouse lines, ES cells, etc. This forms the business of the Expert Group on Improved Resources.

Aim: To enhance access to and promote sharing of information and physical resources for mouse functional genomics research

---

## **Business**

### **Priorities**

The current priorities identified by the committee are:

- Database resources - Identify datasets that are, or will become important over the next 10-15 years.
- Ontologies and standards - Identify ontologies and standards used in the various datasets
- Communication standards - Identify which databases could be integrated and explore methods for doing this
- Biological resources - Determine physical biological resources that are or will need to be maintained
- Sustainability - Consider ways to sustain these physical and information resources beyond the life of the current funding

### **Present**

Members of PRIME resources committee present the meeting in Rome on 2 September 2005.

Paul Schofield, University of Cambridge, UK - Chairman  
John Hancock, MRC Harwell, UK  
Christoph Lengger, GSF, Germany  
Michael Hagn, GSF, Germany  
Jens Hansen, GSF, Germany  
Laurent Vasseur, ICS, France  
Vassilis Aidinis, BSRC Alexander Fleming, Greece  
Claus Nervov, EMBL, Italy  
Werner Müller, GBF, Germany  
Duncan Davidson, MRC HGU, UK  
Eli Reuveni, EMBL, Italy  
Raffaele Matteoni, CNR, Italy  
Elia Stupka, TIGEN, Italy  
Hilary Gates, MRC Harwell, UK - PRIME secretariat  
Mandy Studley, MRC Harwell, UK - PRIME secretariat

---

## Next steps agreed at the meeting

There should be a period of information gathering:

Area	People	Information needed/action
Ontologies	Elia Stupka John Hancock (Duncan Davidson)	1. What ontologies are available and what do we need 2. Technical issues: how can we integrate them, where do they overlap and semantic issues
Communications	Werner Müller John Hancock - phenotyping (Elia Stupka) EBI technician	1. Technical issues: how data can be transferred from one machine to another 2. Standards for data
Database Resources	PRIME office Duncan Davidson Raffaele Matteoni	Information on databases 1. What mouse databases exist 2. The type of database 3. The type of data in them 4. Future information that will need to be stored 5. Relationships between databases 6. Other models and the relationships
Biological Resources	PRIME office	1. What is available currently 2. How can we centralise the information or resource and make it available more widely
Sustainability	Paul Schofield	1. Explore ways to provide funding 2. Meet with Policy Group

### General action

We should produce flyers or leaflet, standard slides and posters for PRIME and the resources committee.

---

## Database resources

We need to identify existing and potential new databases in mouse functional genomics research

- What mouse databases exist; public, restricted and dead
- Potential interaction of databases; including how other model organisms should interact with the mouse
- What issues are important in integrating resources
- What databases we need now and in the future
- What we need to do with databases in order to make them useful, e.g. OMIN, is there any way to convert this to a 'proper' database?

There are two issues to consider:

- The biological requirements for data storage and searches
- The bioinformatics construction of database and searches

It was agreed to set up an information gathering exercise to determine:

- What mouse databases exist
- The type of database
- The type of data in them
- Future information that will need to be stored
- Relationships between databases
- Other models and the relationships between databases

It is important to consider usefulness of the data rather than just focussing on the bioinformatics. This should include assessing the comprehensiveness of the data and identify any gaps.

We should find out what the MFG community needs and produce a map of biological areas that should be covered by databases.

We should talk to other species so as not to duplicate work already being done, e.g. zebrafish, fruitfly, xenopus (XENBASE, gene expression database), chicken, human, ZFIN. This may also give us an idea of the sort of information that might be produced for MFG in the future.

### Improving Search facilities

We should look at improving search facilities.

- Currently genotype links to phenotype information, but not back again.
- It would be good to search on a phenotype and get information on the known genes involved and then be able to get information on other diseases associated with those genes
- Also to search on genotype and get the complex interaction of genes.
- Search on a normal phenotype to get genes involved
- We should consider links with human disease and human clinical databases. The cancer community is a good example of linking between human disease and mouse models. For example, a link could be to describe human disease, link this to MGI to find the mouse disease and give mouse genes and then look for other models.
- Link in with clinical trials to move from mouse model to the disease to the gene and on to clinical trials. (e.g. in the US the CAB and mouse models)
- Gene description with automatic linking to abstracts

- Information on the tools available for each gene would be useful
- Link anatomies with various areas, e.g. expression data
- Better to focus on the sequence rather than the gene. Also have to consider the background that the sequence is on.
- Pathways are useful for describing disease processes. Peter Hunter, in New Zealand, has developed the reactome, a combination of phenotype and pathways.

### **Sources of information on existing databases**

- EU-funded research projects
- Research projects funded by national governments/research funding bodies
- Data-mining
- Web searches

### **Curation**

Curation is clearly a factor in these deliberations. It is labour-intensive to curate information centrally. There is a need for feedback in order to correct mistakes and inaccuracies. Continued curation becomes an issue as funding ends. Manual annotation is too expensive. Automatic curation could be used in some cases. It is important to keep up to date with the successes in automation that could be implemented.

### **Annotation**

Annotation is another time consuming area. Distributed annotation could be used, but it is difficult persuading people to devote their time to this.

There is also a need for quality control of annotation. There should be an ability for people to suggest changes providing they give their authorship details.

### **Submission of data**

There can be a reluctance to submit data to public databases. Including rules on when the data will be made public can help, e.g. remaining restricted for one year after submission. This requires a structure in place to make data public once it has been published. We need to enforce submission of data after publication, e.g. EMAGE data has to be submitted if it is published.

It should be a pre-requisite for publishing that data has been submitted to an appropriate database.

### **Standards for data**

Minimum standards for the data should be set. It is important that the community knows that data has been submitted to standards that meet the standards of the people that want to use the data. We need to consider whether there should be guidelines for databases in new projects.

There may need to be some wet-lab work to aid curation to get some data into a suitable standard to go into the database (cryoarchimics).

Several journals recommend standards, but have not made it a requirement.

We could persuade publishers to impose XML mark-up of the abstracts of published papers. This will get over problems of NOP.

## **Stresses**

There are certain stresses on the system, e.g. for high-throughput imaging where can the images be stored? Should raw data be kept? How best to compare the images? Can they be checked to see how they differ, if there is a discrepancy check them manually to see if they are 'normal', if there is no discrepancy, don't check them.

Sustainability involves good practices and what to do with historical databases.

Could set up an EU working group to explore common issues.

## **Areas for research funding**

There a number of areas that could be funded:

- A research call to work on common standards and information systems
- A Coordination Action for a conference between human and mouse models of disease
- A new project to compare a limited number of diseases, e.g. neurodegeneration and cardiovascular. We may need to develop new techniques within this.

---

## **Ontologies**

### **Ontology working group**

It is proposed to set up a core working group for phenome ontologies. The core workgroup could consist of Janan Eppig, Molly Bogue, John Hancock and Ann-Marie Mallon. Business could be:

- Producing a single web pointing to the data
- Compatibility of current databases
- Working group to cross-map ontologies
- Standards for composition of SOPs
- Possibility of a standard XML format for data

We need to raise the profile of the importance of ontologies as key scientists are asked to give up their time to help develop ontologies. They are willing at first, but slow in their efforts when they realise the magnitude and the amount of their time required.

### **Combining ontologies**

We need to bring other ontologies onboard. It was agreed that the best approach would be to ask key questions by email and then have follow-on meetings in person. These need not be formal meetings, but could be informal discussions in the margins of scientific meetings.

We should make use of Ontoweb and ontology mail-list 'GO-friends' when identifying the ontologies and contacts.

We need to find out how advanced databases and ontologies are and set standards for new.

There are some large dataset that exist and are already integrated. They are not going to change their ontologies overnight.

Jonathan Barnes, EXPAM, cross-species ontology.

With data archiving, there is a need to ensure readability over time. There can be historical data in a database. When an ontology is updated, how can we update databases that have already been annotated?

---

## **Communication standards**

There are two issues:

- Federation of databases
- Maintenance and update of databases

We need input from technicians experienced in linking databases

We need to cross-reference to genome annotation and link to ENSEMBL

There should be a set procedure for announcing updates and versioning.

Are there areas where we should specify minimum standards for data and ontologies?

## **Linking databases**

When cross-linking ontologies, cross-species is almost more important as we are looking for the mouse as a model for human disease.

There can be problems with linking databases an obvious approach would be to link phenome data to gene or sequence data, but there can be phenome data without genome data.

In linking databases we need to recognise that some databases are interesting to a small set of people and will remain separate. There could still be recommended standards and ontologies.

---

## **Biological Resources**

Infrastructures are important to research, but there is often insufficient funding for setting them up or maintaining them. Researchers can be reluctant to submit resources they generate to these repositories.

### **Biological resources**

Examples of biological resources are:

- DNA clones
- Cells
- Mice
- Sperm
- Embryos
- ES cells
- Probes generated in EUREXPRESS

The EUCOMM project will be developing a new resource and database for mouse ES cells and sequence text. This will be developed at the Wellcome Trust Sanger Institute. Dissemination of ES cells and targeting vectors will be performed by the RZPD

There will be a common interface for all clones. There could be overlap with the gene trap database which includes European groups, the US and Japan.

EMMA contains nearly 600 lines, with > 200 more lines in the pipeline

Lines from EUCOMM will also be archived in EMMA.

320 mutant mouse lines from EUCOMM resource which are relevant as disease models will be archived in EMMA and available as frozen embryos for distribution

In addition, EUCOMM aims to establish a central resource for validated Cre zoo lines and plans to produce up to 20 novel inducible Cre lines. All these Cre lines will also be archived in EMMA as frozen embryos and available for distribution.

We need to find ways to continue the funding for EMMA beyond the life of the current EC funding.

### **Linking resources**

The International Mouse Strain Resource (IMSR), hosted by JAX, is a searchable online database of mouse strains and stocks available worldwide, including inbred, mutant, and genetically engineered mice. The goal of the IMSR is to assist the international scientific community in locating and obtaining mouse resources for research. The data content found in the IMSR is as it was supplied by data provider sites. IMSR is also linked with the MGI database providing phenotype data.

In most instances resources should be trans-national rather than federated, that is, they should link existing national resources which may have different structures, rather than impose a common structure

An international effort, the International Federation of Mouse Resources (FIMRe), has been established to explore ways to enhance access to mutant mice for the scientific community by facilitating an easy exchange of preserved material among repositories. Several PRIME partners and EMMA are involved in this exercise.

### **Goals of FIMRE:**

- Coordinate repositories and resource centers to:
  - archive valuable genetically defined mice and ES cell lines being created worldwide
  - meet research demand for these genetically defined mice and ES cell lines
- Establish consistent, highest quality animal health standards in all resource centers
- Provide genetic verification and quality control for genetic background and mutations
- Provide resource training to enhance user ability to utilize cryopreserved resources

Currently EMMA follows a one node per country policy (as requested by the Supervisory Board). This precludes small labs from acting as national satellites of the main national EMMA centers.

We should find out what resources people would like to see distributed, e.g.

- A one-stop shop for mice
- A mouse resources website
- Reagents - a list of ones that work
- RNA
- Tissues

Will we be able to look up a gene and see if an ES cell or mouse is available?

### **Submission of samples/mice to accessible archives**

We could consider funding to transfer resources from small labs to larger stores.

Mouse lines and other resources generated in public-funded projects should be made available to other researchers.

We are left with the problem of finding out what is being produced in these projects and ensuring their submission into archives. One solution would be adding the submission of mice/resources as a written deliverable in individual projects.

---

## **Methods for sustainability**

The issue is not if we want to pay for these resources, but how to. If we do not support them now, we will need to pay to re-create them again. We need to consult with the PRIME Policy Group to explore methods to support these resources. The aim would be to identify the critical components and overcome obstacles.

### **Funding for infrastructures**

There are a number of ways that infrastructures can be funded:

- From grants
- By charge to users
- By commercial partners in partnership allowing them to charge a commercial rate
- By commercial partners contributing towards their cost and becoming privileged users in return
- From core funding at institutes for maintenance of infrastructures
- Do we lobby within the ERA for top-slicing of projects for database infrastructure?

We should look for specific areas where users or commercial funding can sustain these resources. These include curation of ontologies as well as running databases.

A possibility could be to make it part of the research grant to maintain databases and resources or transfer them to another source.

Some new grants may be seeking funding to create a database that already exists. Funding should be conditional on there not already being a database available; co-ordination is needed for this. If there is the funding, should this go to that existing database to provide the required infrastructure. An advisory board would be needed to be able to assess this aspect of funding.

If one large centre takes over the work that a lot of very good small centres are doing already, will this cause problems? Do we want to generate a systems biology/computational biology centre in Europe with oversight of this?

A better alternative might be to keep the smaller centres running and connect them up. Is there a possibility to buy some existing resources to combine them or ensure their continued existence?

## **Actions agreed at the meeting**

### **Information gathering**

There should be a period of information gathering organised through the following working groups:

<b>Area</b>	<b>People</b>	<b>Information needed/action</b>
Ontologies	Elia Stupka John Hancock (Duncan Davidson)	3. What ontologies are available and what do we need 4. Technical issues: how can we integrate them, where do they overlap and semantic issues
Communication	Werner Müller John Hancock - phenotyping (Elia Stupka) EBI Informatician/computer scientist	3. Technical issues: how data can be transferred from one machine to another 4. Standards for data
Database Resources	PRIME office Duncan Davidson Raffaele Matteoni	Information on databases 7. What mouse databases exist 8. The type of database 9. The type of data in them 10. Future information that will need to be stored 11. Relationships between databases 12. Other models and the relationships
Biological Resources	PRIME office Michael Hagn	3. What is available currently 4. How can we centralise the information or resource and make it available more widely
Sustainability	Paul Schofield	3. Explore ways to provide funding 4. Meet with Policy Group

### **General action**

We should produce flyers or leaflets, standard slides and posters for PRIME and the resources committee.

### **Suggestions for co-ordination activities**

1. Ontology sub-group meetings as satellites to major mouse genetics meetings. The idea is to pull people at meeting together for informal discussion. Meetings suggested were CSH, GO, IMGC.
2. Set up a core international working group for phenome ontologies. This could be initiated at an International Phenome meeting set up as a satellite to the Eumorphia third annual meeting in Barcelona in 2006.
3. Can we organise a small formal meeting on these issues at which we carry out some user group consultation? Could this be a deliverable or would it require a specific SSA proposal?
4. The group should present these questions and any preliminary analyses to the Policy Group meeting in 2006.

### **Areas for potential research funding**

The group identified a number of areas for which additional funds might be sought in support of these investigations:

- A research call to work on common standards and information systems
- A Co-ordination Action or SSA for a major conference between human and mouse models of disease
- A new project to compare a limited number of diseases, e.g. neurodegeneration, cancer and cardiovascular for example. We may need to develop new techniques within this.

---

## Glossary

CaBIG	The Cancer Biomedical Informatics Grid is a voluntary network or grid connecting individuals and institutions to enable the sharing of data and tools, creating a World Wide Web of cancer research ( <a href="https://cabig.nci.nih.gov/">https://cabig.nci.nih.gov/</a> )
EMAGE	Edinburgh Mouse Atlas Gene Expression Project <a href="http://genex.hgu.mrc.ac.uk/Emage/database/emageIntro.html">http://genex.hgu.mrc.ac.uk/Emage/database/emageIntro.html</a>
EMPReSS	European Mouse Phenotyping Resource of Standardised Screens ( <a href="http://www.empress.har.mrc.ac.uk/">http://www.empress.har.mrc.ac.uk/</a> )
Ensembl	Ensembl is a joint project between <a href="#">EMBL</a> - <a href="#">European Bioinformatics Institute</a> (EBI) and the <a href="#">Wellcome Trust Sanger Institute</a> (WTSI) to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes ( <a href="http://www.ensembl.org/index.html">http://www.ensembl.org/index.html</a> )
EMMA	European Mutant Mouse Archive ( <a href="http://www.emmanet.org/">http://www.emmanet.org/</a> )
ERA	European Research Area
EUCOMM	European Conditional Mouse Mutagenesis Programme
FIMRe	Federation of International Mouse Resources ( <a href="http://www.fimre.org/">http://www.fimre.org/</a> )
Flybase	FlyBase is a comprehensive database for information on the genetics and molecular biology of <i>Drosophila</i>
IMSR	International Mouse Strain Resource ( <a href="http://www.informatics.jax.org/imsr/index.jsp">http://www.informatics.jax.org/imsr/index.jsp</a> )
XSPAN	XSPAN is a project to support cross-species access to tissue-based genetic information through the development of an internet-based cross-species anatomy network, i.e. a cross-species anatomy ontology integration system. ( <a href="http://www.xspan.org/index.html">http://www.xspan.org/index.html</a> )
GO	The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism ( <a href="http://www.geneontology.org/">http://www.geneontology.org/</a> )
MIAME	Minimum Information About a Microarray Experiment that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment ( <a href="http://www.mged.org/Workgroups/MIAME/miame.html">http://www.mged.org/Workgroups/MIAME/miame.html</a> )
MFG	Mouse functional genomics
MGI	Mouse Genome Informatics provides integrated access to data on the genetics, genomics, and biology of the laboratory mouse. ( <a href="http://www.informatics.jax.org/">http://www.informatics.jax.org/</a> )
OMIM	Online Mendelian Inheritance in Man a database of human genes and genetic disorders ( <a href="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM">http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM</a> )
Ontoweb	A European Union funded project about Ontology-based information exchange for knowledge management and electronic commerce. ( <a href="http://www.ontoweb.org/">http://www.ontoweb.org/</a> )
PRIME	Priorities for Mouse Functional Research across Europe a coordination action funded by the European Commission to identify goals for mouse functional genomics research across Europe ( <a href="http://www.prime-eu.org/">http://www.prime-eu.org/</a> )
SOP	Standard Operating Procedure
XML	Extensible markup language
XENBASE	A database of information pertaining to the cell and developmental biology of the frog, <i>Xenopus</i> ( <a href="http://www.xenbase.org">http://www.xenbase.org</a> )
ZFIN	The Zebrafish Model Organism Database ( <a href="http://zfin.org/zf_info/dbase/db.html">http://zfin.org/zf_info/dbase/db.html</a> )