

EuroPhenome Working Group Meeting

20 June 2006

MRC Mammalian Genetics Unit, Harwell

Attending:

Cornelius Gross	gross@embl.it	EMBL
Eli Reuveni	reuveni@embl.it	EMBL
Christoph Lengger	lengger@gsf.de	GSF
Bruno Simic	bruno.simic@xpertmind.de	GSF
Holger Maier	holger.maier@gsf.de	GSF
Sophie Leblanc	sophie.leblanc@igbmc.u-strasbg.fr	ICS
Laurent Vasseur	vasseur@titus.u-strasbg.fr	ICS
John Wittig	john_wittig@merck.com	Merck
Jacqui White	jkw@sanger.ac.uk	Sanger
Vivek Iyer	vvi@sanger.ac.uk	Sanger
Steve Brown	s.brown@har.mrc.ac.uk	MRC
John Hancock	j.hancock@har.mrc.ac.uk	MRC
Ann-Marie Mallon	a.mallon@har.mrc.ac.uk	MRC
Andy Blake	a.blake@har.mrc.ac.uk	MRC
Alison Walling	a.walling@har.mrc.ac.uk	MRC
Joe Wood	j.wood@har.mrc.ac.uk	MRC
Bob Johnson	r.johnson@har.mrc.ac.uk	MRC
Chris Duran	the.cribsb@gmail.com	MRC
Hilary Gates	h.gates@har.mrc.ac.uk	MRC
Mandy Studley	m.studley@har.mrc.ac.uk	MRC

Chair: John Hancock

Initial presentations

After an introduction and an initial presentation on possible options by Ann-Marie Mallon, we heard presentations on data capture at the four phenotyping centres. Strasbourg and Munich have mature systems. Hinxton and Harwell have systems currently under development. Hinxton's system should be ready by the end of 2006. The timescale for development of Harwell's system is less clear, but they are committed to facilitating data capture for EUMODIC. Because of their proximity to Harwell Bioinformatics, it is likely that their solution will share features with the EUMODIC system.

There was some discussion on whether the existing solutions in use at Strasbourg and Munich could be used for EUMODIC. The Strasbourg system was perceived to be too platform dependent to be portable in this way (as well as being expensive because of licencing conditions and use of Oracle). The Munich system is open source and it is possible that some components may be of use, especially in tracking of lines.

Discussion

The following features of the system to be implemented were discussed:

Central or distributed. There were arguments for using a distributed group of databases but there were a number of arguments in favour of a single database, including: simplicity of data upload, access and mining, the requirement for a distinct project database, the potential to develop a standard for phenotype data storage and the possibility of distributing the entire database as is done, for example, by the Ensembl project.

Data transfer. Given the decision to use a central database, a procedure for data import was the next essential feature. As all centres already have a LIMS module for data capture or are in the process of developing one, the simplest way to do this would be for each centre to generate results files in a standard format which could then either be made accessible via an ftp site or similar, or exported to an import site at Harwell. It was agreed that this file should be encoded using an XML schema. This necessitates the design of a suitable schema. The first important step was therefore to define the parameters measured by each of the EUMODIC SOPs. This could be done initially by studying data files coming out of EUMORPHIA (which shared many of the SOPs to be used by EUMODIC. Dymorphology in particular still poses some problems for data capture because of its descriptive nature, but we hope to draw on experiences from the different centres to deal with this). Conclusions from this exercise would then be shared with the group and the experimental scientists involved in the project to confirm that all the parameters to be measured were covered and in the correct units. It would be essential to gather data on individual mice, to allow co-varying parameters to be identified.

Additional information to be gathered included genotype information and metadata relating to environmental conditions with the potential to affect the results of SOPs.

- Genotype - parameters to be recorded would need to be sex (for individual animals), EUCOMM strain name (as a surrogate for the genetic lesion a particular line represents) and (if necessary) the genetic background (the latter two would be properties of the line as a whole).
- Environmental conditions. A draft list of relevant conditions was compiled as follows:
 - Age/when sampled/test carried out - this could be calculated if dates (or weeks) of sampling and testing were provided in the data set and the week of birth was known.
 - SOP-specific information, e.g. which particular apparatus was used for the experiment. These would need to be derived from a discussion of the SOPs with experimenters to identify which features were potentially variable.
 - Feed - including whether it had been autoclaved
 - Water pH
 - Temperature
 - Humidity (?)
 - Caging - enrichment (type of; there would need to be some discussion of the type of enrichment in use at different centres)
 - Day/light cycle, brightness (lux)
 - No. of air changes in cage
 - Operator of test (this would need to be encoded to identify that different operators had been involved in different data sets without identifying them personally)
 - Health status (screen results - there would need to be some discussion of what this consisted of)
 -

Many of these features would be features of the centre as a whole rather than the individual mouse or line, and could be set up for all data emerging from a given centre. However there was also the possibility that some would change over time, necessitating the possibility of editing the information and date stamping the changes so that appropriate information could be applied to particular data sets.

It was noted that some SOPs produced images, but also that binary data can be passed via XML files and there is the potential to store images as BLOBs in databases or as associated objects.

There was a discussion (which relates to a later discussion about tracking of the pipelines, see below) about when uploaded data should or could be made public. General consensus was that this should be when a data set was complete and there was a strong case for the originating centre to sign off the data set before it was made public. It would be useful for the contributing centres to be able view and, if necessary, re-load data to ensure QC. This in turn would necessitate tracking of version histories via date stamping.

Although ontologies for phenotype representation were not a major theme of the discussion, there was a brief discussion of how data could be annotated with

ontological terms by marking up the SOP with the phenotypic attribute(s) to be measured. It was also suggested that any novel developments in the ontology area be submitted to www.obofoundry.org.

External Access to Data. Design of an appropriate web front end for viewing and analysis of data would be a necessity, but would only become a priority once a solution to the data capture problem was well-developed. Developing this interface could involve all the participants (potentially including some not present at this meeting) and could also take advantage of experience gained in developing the interface to Europhenome. There was a requirement to develop a simple format for reporting the status and phenotypic characteristics of individual lines via the project web pages. This was not discussed in detail, but could involve simple flagging of phenotypic attributes that showed abnormal values in a given line. This would be discussed in more detail at a later meeting. It may be of value to ask secondary phenotyping centres what properties of mouse phenotype would interest them in studying a particular line further. [This discussion to take place at the September 7th/8th meeting in Munich.]

Tracking mice. There was also a requirement for transparency in the process of phenotyping lines. The lines to be phenotyped would therefore need to be presented via the web site, and some means would be needed for individual centres to notify milestone information, such as when a particular line was born and/or went into phenotyping, whether a particular line was homozygous lethal or there was some other problem with the process of deriving the line. Once a cohort was born, there were simple rules governing the progress of mice through the phenotyping pipelines, each SOP being carried out a given number of weeks after birth (i.e. ± 2 days of a specified date) and the complete process being finished by 14 weeks. Data should be available on the database within a specified period after this 14 week period was over, with a target submission deadline of 4-6 weeks after completion. (This period remains to be agreed by the phenotyping centres.)

It was important that each mouse be issued with a unique EUMODIC ID that would allow linking of their data. This ID would not need to be known to the centres, but the centres would need to associate two IDs with their data: their local mouse ID (which could be made unique within EUMODIC by appending a centre code) and their EUCOMM strain name. There would be a particular need to track UIDs when mice went into secondary phenotyping at additional centres. [There is also a need to define SOP parameters coming out of secondary screens, but this was not a priority at present.]

Process

As well as the two follow-up meetings scheduled for September and December, it was agreed that we would gather a smaller working group to look at the parameters to be measured and, more generally, at the XML specification. This working group would preferably consist of one representative per centre and be drawn from:

- Sanger: Dave Melvin, Niels Adams
- GSF: Christoph Lengger, Holger Maier
- ICS: Laurent Vasseur, Sophie Leblanc

- Harwell:
 - Informatics: Ann-Marie Mallon, Simon Greenaway, Chris Duran
 - Data capture: Joe Wood
 - Monterotondo: Eli Reuveni, Cornelius Gross
 - Secretariat: Hilary Gates
- although this list is not set in stone.

The series of meetings would be augmented by telephone conferences to be held at appropriate intervals to ensure that the process maintains momentum. A first meeting will be scheduled for the end of July 2006.

Main Points

In summary, it was agreed that:

- There would be a single database that would hold all the data emerging from the EUMODIC project
- Data would be submitted to the database via a custom XML to be agreed by a working group representing the main phenotyping centres and the EUMODIC scientists in general
- The major phenotyping centres commit to having the capability to generate the requisite XML files from their data capture environments by the start of the project and preferably by the beginning of 2007.

A working prototype of the data capture system should be available by the beginning of 2007 and a working version by the end of June 2007.