# IPAD-MD

# Data and Resources Expert Group Workshop on

# Improving curation and annotation tooling for mouse models of rare disease

## (D7.2)

## 08-09 December 2016

## Freising, Germany

# MEETING MINUTES

## 1 Agenda

December 8th

**13:00 – 15:30** / Session 1 / **Current curation practices**

- Introductions
- MGI Curation process (Susan Bello)
- IMPC automated curation (Terry Meehan)
- BioMedBridges pilot on diabetes (Nathalie Conte)
- Current annotation of strains at CNR (Raffaele Matteoni)

**15:30 – 16:00** / Coffee break

**16:00 – 18:00** / Session 2 / **Discussion: The semantic resources and tools**

- Review as a group and discuss
    - Phenotype Ontologies (MP, MPATH, HPO)
    - Disease Vocabularies (OMIM,ORPHANET, Disease Ontology, EFO, NCI-T, others)
    - Curation Tools (OLS, Bioportal, ZOOMA, CEDAR)
- What does the future hold for curation of disease models?
    - Improved standardisation (ie. MUNDO)
    - Phenopackets
    - IMPC: next five years
    - MOD in the cloud

**19:30** / Dinner / Restaurant "Die Molkerei", Marriott Hotel

December 9th

**09:00 – 10:30** / Session 3 / **How can we collaborate?**

- Should/How can coordinate manual annotations of disease models?
- How best to use automated tools
- Data exchange formats

**10:30 – 11:00** / Coffee break

**11:00 – 12:00** / Session 4 / **Action items**

- Review items discussed
- Action items matrix
- Prioritise based on need vs difficulty

## 2 List of participants

- Marzia Massimi — CNR Monterotondo, Rome
- Raffaele Matteoni — CNR Monterotondo, Rome
- Nathalie Conte — EBI, Hinxton
- Terrence Meehan — EBI, Hinxton
- Christine Schütt — HMGU, Munich
- Michael Hagn — HMGU, Munich
- Patricia da Silva Buttkus — HMGU, Munich
- Manoj Chinnasamy — INFRAFRONTIER GmbH, Munich
- Michael Raess — INFRAFRONTIER GmbH, Munich
- Philipp Gormanns — INFRAFRONTIER GmbH, Munich
- Sabine Fessele — INFRAFRONTIER GmbH, Munich
- Susan Bello — JAX, Bar Harbor

## 3 Minutes

### 3.1 Minute takers

Sabine Fessele and Terrence Meehan

### 3.2 Aims of the workshop

1. Facilitate communication between curators that have an interest in annotating animal models of human disease
2. Discuss the strengths and weaknesses of curatorial resources (ontologies, tools, etc.)
3. Investigate synergies between groups to accelerate annotation of rare disease models and their uses

Advances in genome editing and other technologies are accelerating the generation and characterization of model organisms. Manual curation of disease models is important but challenging giving the accelerated pace of discovery. New automated methods to identify rare disease models show promise but lack the expert knowledge curation brings. In this

workshop, we will bring together curators to discuss their annotation pipeline, the tools they use and investigate sharing of resources to improve the curation process.

### 3.3 Current curation practices

The meeting was started by informing each other about the curation processes in the different organisations or projects.

Sue explained MGI's rules for disease annotation, which include the principles that annotations are only inferred if the mutation is causative (as stated in the publication), that positives trump negatives (if one publication says something is a model, but another says it is not a model, it will be annotated as a model) and that 'NOT model' annotations are made if there was a reason to expect it, but it was proven negative. Current topics at MGI are the upcoming switch from using Online Mendelian Inheritance in Man (OMIM) to Disease Ontology (DO) for disease annotation. MGI has started splitting DO terms (as DO combines OMIM terms), will first switch their front-end display, but keep up using OMIM curation-wise. The MouseMine data warehouse system will provide OMIM and DO mappings in parallel. In the long term OMIM annotation will be discontinued. MGI IDs will stay. (See 3.5 below for related information.)

Terry gave an overview of the IMPC disease annotation pipeline and showed examples of some of the novel disease models and disease gene candidates they have uncovered. The PhenoDigm phenotypic similarity algorithm allows IMPC to identify good mouse models for known, as well as candidates for novel disease-gene associations and suggestions for secondary phenotyping projects. Translational applications using IMPC data in large scale sequencing projects were also introduced (e.g. application of the Exomizer tool for the NIH Undiagnosed Diseases Program).

Nathalie presented the DIAB ontology, which is an ontological representation of expert knowledge about type 2 diabetes. It was developed based on text mining and expert review. In the meantime this pragmatic and rapid method has also been used for other areas at EBI (e.g. inflammatory bowel disease), and takes about 6 months for a disease field.

Raffaele described which manual and computational methods are applied by the INFRAFRONTIER curators at CNR to maintain official strain nomenclature and annotate EMMA strain records. The process involves proposing new allele/gene records to MGI, e.g.

for unpublished targeted/transgenic strains or spontaneous mutants that have not been registered by MGI, yet. Preparing for the new INFRAFRONTIER2020 task of also annotating EMMA strains with phenotypic and disease model data, INFRAFRONTIER has conducted a preliminary screening of manually curated EMMA strain records. After doing the necessary validation steps (including matching allelic composition and genetic backgrounds), INFRAFRONTIER intends to display these annotations on the public EMMA strain detail pages and highlight some main disease areas.

### 3.4 Discussion on semantic resources and tools

The group reviewed available phenotype ontologies, disease vocabularies and curation tools and how they can best be used for disease model curation and annotation.

It was noted that, as the Human Phenotype Ontology (HPO) only uses rare disease terms, there is no ontology for common diseases. The big divide between bioinformatics and medical informatics represents another major challenge. It was highlighted that trait ontologies (e.g. the Vertebrate Trait Ontology) differ from phenotype ontologies by just including things that are measured in an assay without giving a quality measure (e.g. ‘reduced’, ‘increased’, ‘abnormal’).

The pros and cons of disease vocabularies and their license models were briefly discussed.

Nathalie gave a demonstration of the Zooma tool, which was developed at EBI for mapping free text to ontology terms based on a curated repository of annotation knowledge. The Zooma process includes feedback loops and therefore repeated rounds will lead to better results.

### 3.5 Curation of disease models – the future

Sue updated the group about the activities of the **Alliance of Genome Resources (AGR)** group that aims to unify the six model organism databases (MODs). A merged model organism database is assumed to be more sustainable and to make it easier for researchers, as they will only have to search a single site. (Note: Existing Mines based on the InterMine platform (e.g. MouseMine) can already offer cross-organism queries, but these are not easy for end-users.)

Same as MGI, IMPC and INFRAFRONTIER, the AGR sees the value of providing researchers access to high-quality expertly curated information, which has a high impact on biomedical research.

Previously disease annotations were done to different entities (genes, allele, strains, …) in the different MODs. In the future the gene will be the only cross-model object and all MODs will use DO. In line with this MGI will soon release their new Disease Ontology Browser. The Human - Mouse: Disease Connection (HMDC) query will also be changed and MGI expects that this will broaden their user base.

Terry mentioned that **IMPC** can handle **aging** phenotypes, but that it still needs to be decided how to incorporate them into disease model annotations. It was suggested to have a look at the Rat Genome Database (RGD), as they have already implemented this. Related to the IMPC aging pipeline it was further suggested to make use of the Mouse Phenome Database (MPD), as it gives access to baseline data over time.

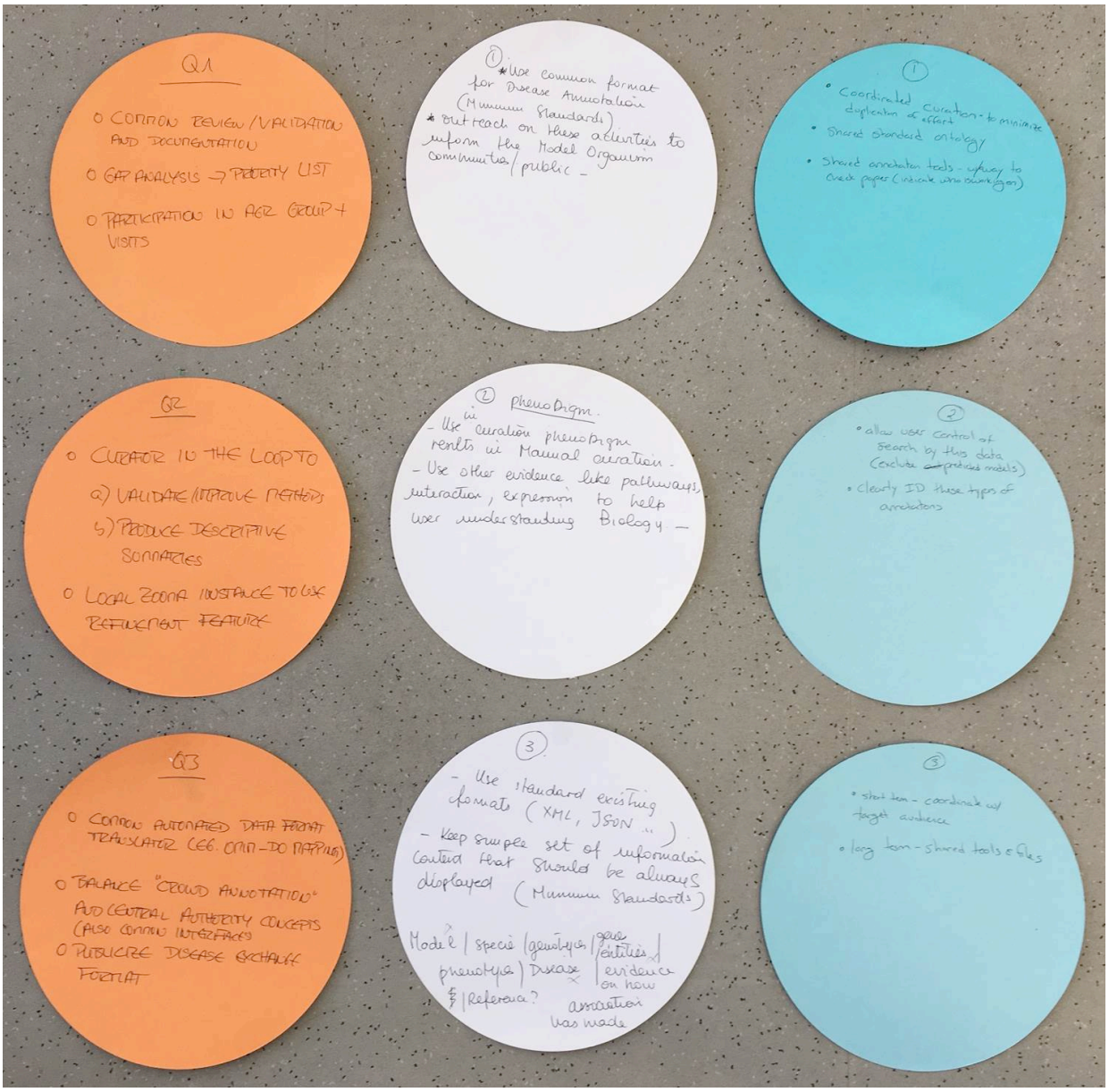
IMPC lines up with **precision medicine** initiatives like Genomics England. IMPC's role will be to do Exomiser prioritization including IMPC data, and Genomiser analysis on negative cases. Variants from human sequences will be sent though the IMPC pipeline (CRISPR point mutations).

Terry presented a slide deck on **PhenoPackets** that was provided by Melissa Haendel. The PhenoPacket concept was developed to define a standard exchange format for phenotypes that does not just describe the entities, but also the relationships between them. The PhenoPackets community envisions everyone (scientists, clinicians and laypersons) to use the new Phenotype Exchange Format (PXF). To allow this the AGR will try to put in layperson synonyms for annotation by them. Journals could act as gatekeepers for registering PhenoPackets and if everything is available in the same format this could facilitate identification of conflicting evidence and patient matchmaking.

Since it is not defined what goes into a PhenoPacket and since it does not have a unique identifier - you can use the concept to annotate to any object you would like to make your annotation to – MGI would package the complete genotype – phenotype set, rather than to just annotate to gene only. The question for application by IMPC would be when to generate a PhenoPacket, i.e. several times or only when phenotyping for a strain is complete.

## 3.6 Collaboration

To develop ideas on how best to coordinate annotation of disease models between INFRAFRONTIER and MGI a brainstorming session in small groups was held. From the results that are captured in the picture below a number of decisions and action items were derived (see 3.7)



## 3.7 Action items

| Topic/Decision | Action item | Responsible (Timeline) |
|---|---|---|
| Curation process documentation is available in MGI Wiki. This is private, but can be further shared with CNR (also already shared with Monarch Initiative). | Share curatorial information with CNR | Sue<br>(Dec 2016) |
| Review annotation process after reading documentation | Phone call | Marzia, Sue<br>(Jan 2017) |
| Initial disease and phenotype information format | Google spreadsheet to define minimal information | Sue<br>(Jan 2017) |
| Produce list of INFRAFRONTIER/EMMA strains with mutations orthologous to human genes associated with Mendelian diseases in OMIM and Orphanet.<br>Subtract from list mice already declared disease models by MGI (potential training set). | Send list to Sue | Marzia, Raffaele<br>(Spring 2017) |
| Annotate strains on INFRAFRONTIER priority list (output from point above)<br>Process<br>• MGI sheds papers to CNR to work on disease annotation (phenotype annotation required first)<br>• CNR provides annotations<br>• MGI pastes in annotations (less effort than reading the paper and writing down data)<br>• Phone calls to review annotations | Regular process to be established between MGI and CNR | MGI: Sue<br>CNR: Marzia |
| Options to be evaluated for face to face meeting<br>• CNR visit to JAX to be trained<br>• Biocuration meeting in March in California<br>• Organize satellite to INFRAFRONTIER2020 user meeting in Greece as “Disease model annotation for 21st century – DO workshop”<br>• Anna Anagnostopoulos as Europe-based MGI curation contact | | Sue, Marzia<br>Raffaele, Marzia<br>Terry, Sabine |

| | | |
|---|---|---|
| Use INFRAFRONTIER/CNR as pilot user for web based annotation tool in the future | Use case | |
| Alliance of Genome Resources (AGR): check if CNR could sit in on disease and phenotype working group call and/or receive meeting notes | Check with Cynthia | Sue |
| PhenoDigm: consensus curator in the loop would be helpful for validating and improving algorithm | Send INFRAFRONTIER list (from above to Damian and Terry) | Raffaele |
| Sue's paper on rules for associating disease to gene (as opposed to genotype) (numbers of cre conditional genotypes vs simple genotypes) | Read paper | All |
| Use MousMine to pull more complete disease and phenotype associations and transgene models. | Get information from MouseMine | Philipp |
| Improvements to DO<br>• MGI and RGD are working with DO to expand coverage of OMIM and improve structure with relationships (Elvira is working on this)<br>• Sue reporting to AGR, DO | Find notes from IMPC workshop and send to Sue | Terry |
| DO changes propagated to live file | Check OLS to see how relationships are displayed | Nathalie |
| Improving ZOOMA for ontologies and portability | Follow up on MP functionality | EBI |
| Prepare INFRAFRONTIER for switch over to MGI curation of DO | | Philipp (no timeline yet) |

## 3.7 Further documentation

Presentations

- MGI-disease-model-curation_Sue-Bello.pptx
- Phenodigm-and-IMPC_Terry-Meehan.pptx
- Ontologies-and-tools_Nathalie-Conte.pdf
- INFRAFRONTIER-EMMA-strain-annotation_Raffaele-Matteoni.pptx

- Phenopackets_Terry-Meehan-on-behalf-of-Melissa-Haendel.pptx
- IMPC-future_Terry-Meehan.pptx

MGI documentation

- cur_Allele disease module - MGI_WIKI.pdf
- cur_Genetic background annotation - MGI_WIKI.pdf
- cur_Phenotype Annotation - MGI_WIKI.pdf

Available for download on the internal INFRAFRONTIER webpage at https://www.infrafrontier.eu/internal/ipad-md-project