Mamm Genome (2009) 20:457–461
DOI 10.1007/s00335-009-9208-3

COMMENTARY

Mouse, man, and meaning: bridging the semantics of mouse phenotype and human disease

John M. Hancock · Ann-Marie Mallon · Tim Beck · Georgios V. Gkoutos · Chris Mungall · Paul N. Schofield

Received: 15 April 2009 / Accepted: 10 July 2009 / Published online: 2 August 2009

Abstract Now that the laboratory mouse genome is sequenced and the annotation of its gene content is improving, the next major challenge is the annotation of the phenotypic associations of mouse genes. This requires the development of systematic phenotyping pipelines that use standardized phenotyping procedures which allow comparison across laboratories. It also requires the development of a sophisticated informatics infrastructure for the description and interchange of phenotype data. Here we focus on the current state of the art in the description of data produced by systematic phenotyping approaches using ontologies, in particular, the EQ (Entity-Quality) approach, and what developments are required to facilitate the linking of phenotypic descriptions of mutant mice to human diseases.

The laboratory mouse is a pivotal organism in understanding mammalian biology and gaining insight into human diseases. Now that the mouse genome sequence is available and annotation of genes and regulatory sequences within it is improving rapidly, attention is shifting to how genome information can be used to better understand mammalian biology (Brown et al. 2006). Broadly, this knowledge can be applied in two ways: (1) to gaining a systems-level understanding of mouse biology leading to an understanding of the effects of mutations and other interventions on a variety of pathways giving rise to mutant phenotypes, and (2) to the direct identification of mouse mutants with features that map directly onto human disease phenotypes or aspects thereof. These new mouse strains can then serve as disease models to improve our understanding of human pathobiology and support the development of new approaches to therapy.

Both of these grand projects, which overlap to a considerable extent, demand the systematic acquisition of extensive phenotype data on mouse strains. For the systems-level approach, this is needed to relate phenotypes that may have a common cause but different ultimate manifestations at the level of the organism; for the more directed approach it is important to identify as many contributory phenodeviant components of the overall phenotype (endophenotypes) as possible to ensure that the phenotype of a given strain can be accurately linked to its cognate human disorder.

Mirroring the high-throughput vision of biology pioneered by the genome projects, mouse genetics is moving into an era of large-scale data-gathering on phenotypes based on the ability to carry out large-scale genome manipulation, sequencing, and phenotyping. The systematic collection of mouse phenotype data started with the Mouse Phenome Project (Bogue and Grubb 2004), initiated in 1999, which collects data on a large panel of inbred lines. A database, the Mouse Phenome Database (MPD) (Bogue et al. 2007; Grubb et al. 2004), is an integral part of

J. M. Hancock (✉) · A.-M. Mallon · T. Beck
Bioinformatics Group, MRC Harwell, Harwell, Oxfordshire
OX11 0RD, UK
e-mail: j.hancock@har.mrc.ac.uk

G. V. Gkoutos
Department of Genetics, University of Cambridge, Downing
Street, Cambridge CB2 3EH, UK

C. Mungall
Lawrence Berkeley National Laboratory, Berkeley, CA 94720,
USA

P. N. Schofield
Department of Physiology, Development and Neuroscience,
University of Cambridge, Downing Street, Cambridge
CB2 3DY, UK

this project. The MPD serves to integrate data captured at a variety of centres both to allow for data dissemination and to provide tools for the comparative analysis of strains.

The Mouse Phenome Project collects data on many mouse lines in a single laboratory. For example, data from a large-scale aging project carried out at The Jackson Laboratory (http://agingmice.jax.org/index.html) will soon deposit age-related phenotype data for 32 inbred strains of mice into the MPD. This approach to phenotyping ensures reproducibility of individual types of measurement but has problems of scalability—how many lines is an individual laboratory able to analyse?—and confounding factors such as environment, assay, or equipment variation mean that any meta- or coanalysis of different measurements is not possible across laboratories. The problem of scalability is addressed by an alternative structure, pioneered by the EUMORPHIA project in Europe (Brown et al. 2006), in which more than one centre carries out all phenotyping tests, and efforts are made to make all of those tests as reproducible as possible between centres. EUMORPHIA's aim was to produce a panel of reproducible phenotyping tests that could be used in such a project in the form of SOPs (standard operating procedures) and to test these on a number of inbred lines. The project resulted in the EMPReSS collection of SOPs (Brown et al. 2005). Comparability between tests is also addressed by this approach, although it remains to be established to what extent comparison of lines between laboratories can be achieved.

A natural complement to the Mouse Phenome Project and EUMORPHIA is the collection of phenotype data on mutant mouse lines, rather than inbred background strains, using the same concept of standardized phenotyping procedures. This is currently the aim of the EUMODIC project (http://www.eumodic.org/), a follow-on from EUMORPHIA. Here, four large mouse clinics—MRC Harwell and the Sanger Institute in the UK, the Helmholtz Zentrum in Germany, and the Institut Clinique de la Souris in France—carry out phenotyping using standard pipelines, making use primarily of gene knockout lines from the EUCOMM project (Friedel et al. 2007), part of the International Knockout Mouse Consortium (Gondo 2008). The benefit of using EUCOMM lines in EUMODIC is that because all mice are derived from the same embryonic stem (ES) cell line, there is minimal interindividual genetic variation, allowing for more robust identification of phenotype-gene relationships.

Both EUMORPHIA and EUMODIC are distributed projects with data being generated at different international centres. The only sensible approach to such data-gathering exercises, as illustrated by the Mouse Phenome Project, is to set up a central data repository. The SOP collection from EUMORPHIA was collected in the EMPReSS database (Green et al. 2005) (http://empress.har.mrc.ac.uk/), which contains all the EUMORPHIA SOPs in a structured form. Data from EUMORPHIA, and subsequently EUMODIC, have been collected in the EuroPhenome database (Mallon et al. 2008) (http://www.europhenome.org).

Of course, these high-throughput approaches are of value only if complemented by the gathering of more detailed information on lines of particular relevance to disease. Pathological examination, which is not part of high-throughput phenotyping pipelines because of its cost and relatively time-consuming nature, is an important part of this because without this the lines that are identified are less likely to be adopted by the relevant disease communities.

Inspection of the data generated by the EMPReSS SOPs and held in EuroPhenome immediately reveals that phenotyping data are much more diverse than the kinds of data bioinformaticians are used to handling, e.g., sequence data. Many of the observations that need to be held in a phenotyping database are either descriptive or involve inferences from raw data. Any type of data that to a significant extent comprises free text causes immediate problems for data analysis because different experimenters may use the same term for different things or different terms for the same thing. This semantic problem has led to the increasingly broad uptake in biology of ontologies, starting with the gene ontology (Ashburner et al. 2000). Ontologies are formal structures that consist of two elements: standard terms that can be used to describe a domain of knowledge and relationships linking those terms. Classically, in bio-ontologies these relationships are of the form "is_a," e.g., an eye "is_a" sensory organ, or "part_of," e.g., an eye is "part_of" the head. The benefit of including such relationships is that it is possible computationally to relate data on the eye, in this case, to other data relating to the head or to other sensory organs.

The first use of ontologies to describe mouse phenotypes was the Mammalian Phenotype Ontology (MP), which continues to be developed at The Jackson Laboratory (Smith et al. 2005). This is used in the Mouse Genome Database (MGD) (Blake et al. 2009) to annotate the abnormal phenotypes of mouse strains and lines. The MP currently contains over 9000 terms and is an immensely powerful tool for mouse line annotation. However, it is not suitable for the detailed description of data generated by high-throughput phenotyping projects for a number of reasons. First, all phenotypes described by the MP are abnormal, whereas much, perhaps most, of the data generated by high-throughput phenotyping is normal. Indeed, within EuroPhenome no judgment of normal or abnormal is made on raw data held in the database; any such annotations are made by inference, which is increasingly done by automatic reasoning. Second, MP does not allow the description of quantitative values obtained in phenotyping experiments.

The MP is described as a "pre-composed ontology" because its terms contain combinations of terms that exist in other, more fundamental ontologies such as the mouse anatomy ontology and the Gene Ontology. An alternative to the pre-composed structure of the MP ontology is the "post-composed" structure implemented in the EQ approach (Gkoutos et al. 2004, 2005). The EQ approach is based on the Phenotype And Trait Ontology (PATO), an ontology of phenotypic qualities intended for use in a number of applications, primarily phenotype annotation. According to this approach, phenotypic descriptions can be abstracted into two parts: an *entity* that is affected (the thing for which measurements are made), be it an enzyme, an anatomical structure, or a biological process, and the *quality* of that entity, described either qualitatively or quantitatively. At a bare minimum all phenotypes are described using a class expression consisting of a quality class (from PATO) differentiated by a bearer entity class (from some other open biomedical ontologies, OBO ontology) using the *inheres_in* relation (from the OBO Relation Ontology RO). One way of expressing phenotype annotation using PATO is the so-called pheno-syntax, which is adopted in EuroPhenome. In its simplest form a pheno-syntax tuple can be E = MA:liver Q = PATO:hyperplastic to describe a "hyperplastic liver" phenotype. (Formally, this description is compiled using the IDs of terms from the ontologies rather than their names because names are not necessarily fixed, whereas IDs are stable; we are using term names here for clarity.)

The pre- and post-composition approaches are completely compatible provided that equivalence relationships to EQ descriptions are generated for pre-composed ontology terms. The MP, for example, does have a term describing "liver hyperplasia" (MP:0005141). If an equivalence relationship mapping is provided, then these two descriptions can be used interchangeably. In OBO 1.2 syntax, this could be represented as

```
[Term]
id: MP:0005141 ! liver hyperplasia
intersection_of: PATO:0000644 ! hyperplastic
intersection_of: inheres_in MA:0000358 ! liver
```

This is a very powerful aspect of EQ because it provides the potential to link different ontologies at a basic level, allowing for a common mapping of different phenotype ontologies through logical definitions and "phenotype (EQ) statements." The liver hyperplasia example is a relatively simple one which serves to illustrate how EQ statements can be constructed. However, many disease terms are much more complex and require a more detailed description. EQ allows for complex phenotype descriptions such as disease terms by allowing the creation of a phenotypic profile formed by several EQ statements. For example, for the term "osteoporosis" a set of EQ statements would be required to describe "increased bone resorption" and "decreased osteogenesis" that *results_in* "decreased bone mass" and "increased bone fragility."

MP terms, which have been created largely by phenotyping scientists using community-agreed-on terminology, can be standardised by using EQ-based logical definitions. For example, the MP term *belly spot* (MP:0000373), which is a term widely used by the phenotyping community, defines the pigmentation phenotype characterised by "the appearance of a round area of white fur on the belly" (this is the MP definition). The *belly spot* term could be logically defined based on the EQ approach as "white" (PATO:0000323) "spotted" (PATO:0000333) "coat hair" that is part_of "abdomen" (MA:0000029).

Such an EQ-based logical definition does not define a phenotype as being intrinsically "abnormal" and can provide new relationships between the MP terms through the use of predefined entity and quality terms. For example, the MP uses the labels "belly" and "abdomen" to define the same concept, in line with community use. However, MA defines the single concept (abdomen) which can be used to link the MP terms containing both "belly" and "abdomen" at the EQ level.

These kinds of mappings are a new area of development in the phenotype ontology field and require some significant issues to be addressed. First, there is a need to appropriately logically define the relevant pre-composed ontologies. This is an issue of both manpower and expertise as many of the terms in pre-composed ontologies have specialist meanings that require specialist input into the definition process to ensure that they are broken down correctly. Second, the underlying ontological infrastructure is still patchy. While anatomy and quality ontologies exist for mouse phenotypes, there is currently no behaviour ontology that can be used to define behaviour-related phenotype terms. From the perspective of human disease there is a need for an intermediate level between disease terms and human "phenotypes" that can be linked to mouse phenotypes (Schofield et al. 2008). This missing layer may be filled by the new Human Phenotype Ontology (HPO) (Robinson et al. 2008), which currently describes phenotypes that are components of diseases in the OMIM database, but this is still in the early stage of development and the provision of logical definitions for HPO and mapping to the mouse will be required.

A long-term aim of building these links is to facilitate the construction of reasoner software that can link phenotypes in the mouse automatically to human diseases and phenotypes in other organisms. EQ can already be used in the automatic annotation of phenotypes in mutant lines (Beck et al. 2009). Annotation of lines in EuroPhenome, which is currently the only mammalian database to use the EQ formalism, to a considerable extent is automated and takes advantage of the close relationship between the EMPReSS SOP database and the data on baseline strains and mutant lines in EuroPhenome. EMPReSS SOPs define the traits (or parameters) to be measured, e.g., "body weight." The raw body weight data for the mutant line and the corresponding background strain are stored in EuroPhenome. If the mutant line's body weight is found to be significantly increased or decreased after the application of a statistical test, then the relevant EQ annotation is applied to the mutant strain, e.g., "adult mouse" (MA:0002405) "increased weight" (PATO:0000582). For more detail on the automated annotation in EuroPhenome see Beck et al. (2009).

An important area for future development in phenotype bioinformatics will be the linking of disparate data sets to form a phenotype semantic web (Gkoutos 2006). As with all semantic web applications, this will depend on standard formats for semantic representation (ontologies) and data transfer (XML and RDF). From the perspective of ontologies it will be essential to ensure that all the necessary definitions and cross-mappings are established and maintained. At the level of data transfer, data standards for describing phenotype data still need to be established, although formats such as the XML used to transfer data within EUMODIC represent a first step in this direction. Encouragingly, there is broad agreement in the mouse phenotype database community (Mouse Phenotype Database Integration Consortium 2007), and more broadly in the mouse functional genomics community (Hancock et al. 2008), that these developments are needed and the Interphenome Consortium (Mouse Phenotype Database Integration Consortium 2007) has met twice a year since 2007 to work toward these goals. It is possible to link many databases together at the naïve level using tools such as BioMart (Smedley et al. 2005), and more complex integration of data using a variety of mechanisms has recently been demonstrated by the European CASIMIR consortium (Smedley et al. 2008). However, a fully functional integration of mouse phenotype data ultimately will require a more sophisticated, flexible infrastructure, mostly likely based on web services (Stein 2002, 2008). A survey of mouse functional genomics databases (Hancock et al. 2008), also by CASIMIR, suggested that these more sophisticated technologies are being taken up with increasing enthusiasm by database providers, raising the prospect that this cyberinfrastructure will come into existence over the next few years.

Acknowledgments JMH and AMM thank the UK Medical Research Council for support. TB is supported by EUMODIC, funded by the European Commission under contract number LSHG-CT-2006-037188. GVG thanks the UK Biotechnology and Biological Sciences Research Council for support. CASIMIR is funded under Framework Programme 6 of the European Commission (LSHG-CT-2006-037811).

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25:25–29

Beck T, Morgan H, Blake A, Wells S, Hancock JM et al (2009) Practical application of ontologies to annotate and analyse large scale raw mouse phenotype data. BMC Bioinformatics 10(Suppl 5):S2

Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE (2009) The Mouse Genome Database genotypes::phenotypes. Nucleic Acids Res 37:D712–D719

Bogue MA, Grubb SC (2004) The mouse phenome project. Genetica 122:71–74

Bogue MA, Grubb SC, Maddatu TP, Bult CJ (2007) Mouse phenome database (MPD). Nucleic Acids Res 35:D643–D649

Brown SDM, Chambon P, Hrabé de Angelis M, EUMORPHIA Consortium (2005) EMPReSS: standardised phenotype screens for functional annotation of the mouse genome. Nat Genet 37:1155

Brown SD, Hancock JM, Gates H (2006) Understanding mammalian genetic systems: the challenge of phenotyping in the mouse. PLoS Genet 2:e118

Friedel RH, Seisenberger C, Kaloff C, Wurst W (2007) EUCOMM—the European conditional mouse mutagenesis program. Brief Funct Genomic Proteomic 6:180–185

Gkoutos GV (2006) Towards a phenotypic semantic web. Curr Bioinformatics 1:235–246

Gkoutos GV, Green EC, Mallon AM, Hancock JM, Davidson D (2004) Building mouse phenotype ontologies. Pacific Symp Biocomput 178-189

Gkoutos GV, Green EC, Mallon AM, Hancock JM, Davidson D (2005) Using ontologies to describe mouse phenotypes. Genome Biol 6:R8

Gondo Y (2008) Trends in large-scale mouse mutagenesis: from genetics to functional genomics. Nat Rev Genet 9:803–810

Green EC, Gkoutos GV, Lad HV, Blake A, Weekes J et al (2005) EMPReSS: European mouse phenotyping resource for standardized screens. Bioinformatics 21:2930–2931

Grubb SC, Churchill GA, Bogue MA (2004) A collaborative database of inbred mouse strain characteristics. Bioinformatics 20:2857–2859

Hancock JM, Schofield PN, Chandras C, Zouberakis M, Aidinis V et al (2008) CASIMIR: coordination and sustainability of international mouse informatics resources. 8th IEEE international conference on bioinformatics and bioengineering. Athens. Greece, IEEE, pp 382–387

Mallon AM, Blake A, Hancock JM (2008) EuroPhenome and EMPReSS: online mouse phenotyping resource. Nucleic Acids Res 36:D715–D718

Mouse Phenotype Database Integration Consortium (2007) Integration of mouse phenome data resources. Mamm Genome 18:157–163

Robinson PN, Kohler S, Bauer S, Seelow D, Horn D et al (2008) The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 83:610–615

Schofield PN, Rozell B, Gkoutos GV (2008) Towards a disease ontology. In: Burger A, Davidson D, Baldock R (eds) Anatomy ontologies for bioinformatics: principles and practice. Springer-Verlag, London, pp 119–132

Smedley D, Swertz MA, Wolstencroft K, Proctor G, Zouberakis M et al (2008) Solutions for data integration in functional genomics: a critical assessment and case study. Brief Bioinform 9:532–544

Smedley D, Haider S, Ballester B, Holland R, London D et al (2005) BioMart - biological queries made easy. BMC Genomics 10:22

Smith CL, Goldsmith CA, Eppig JT (2005) The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol 6:R7

Stein L (2002) Creating a bioinformatics nation. Nature 417:119–120

Stein LD (2008) Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. Nat Rev Genet 9:678–688