# One Medicine: Integrating mouse and human disease phenotypes

Paul N. Schofield, Georgios V. Gkoutos, John Sundberg, John M. Hancock , The CASIMIR Consortium

***Abstract*—One of the major challenges of the post-genomic era is the coding and integration of phenotype data from humans and model organisms such as the mouse. This is required for phenotype-driven gene function discovery and to maximally leverage the power of comparative pathobiology. In this paper we review the current state-of-the-art for phenotype and disease description in mice and humans and discuss ways in which a common approach to phenotype description will allow the bridging of the gap between the two species.**

## I. Challenges of the post-genomic era

As we enter the post-genomic era, the model-organism approach to the understanding of fundamental biological processes and disease has come of age. With the completion of the mouse and human genomes and the development of genetic toolkits enabling reverse genetic approaches, the challenges of functional genomics have become predominantly those of phenotyping, using and integrating the huge volume of complex data now emerging, and developing a shared and sustainable infrastructure to integrate and exploit data being generated around the world. Biology has become "Big Science" in the way that high energy physics and astronomy have been traditionally characterised. With this comes the need to provide a level of infrastructure and co-ordination not previously considered, at least on this scale.

In response to these needs the European Commission has funded the CASIMIR coordination action (http://www.casimir.org.uk, and the paper in this volume by Hancock and others [1] to examine the current state of the art in mouse functional genomics, to identify current strengths and weaknesses in the infrastructures and standards used by the community and to recommend a sustainable framework for the sharing and use of data. (For discussion of other aspects of the CASIMIR remit see also papers [2], [3] in this volume).

Integration of phenotype data from disparate sources and organisms presents one of the most difficult tasks faced in functional genomics, but one of the most essential. The concept of *phenotype* is used in a variety of ways, not all completely compatible; for example descriptions of clinical diseases (signs and symptoms), pathological lesions and entities, summative disease nomenclature (e.g. syndromes), appearance of mutants and strains, genetically determined traits of strains, and at the lowest level transcriptome and gene expression patterns. Properly defined the *phenome* constitutes the sum total of the genetically determined traits as manifested under the prevailing environmental conditions and a *phenotype* is an observable property of the organism under those circumstances. This means, for example, that development of a *tumour* is a *phenotype*, a measurable manifestation of a heritable trait that might be described as *tumour frequency* or *lifetime tumour risk* for a particular tumour type. However the term *phenotype* is often used as a proxy for a trait, such as an *heritable predisposition*, especially in the description of human disease; the resulting confusion, which is an especial problem for complex traits, is discussed below.

In the first half of the CASIMIR project we have been gathering information on existing description formalisms, mainly ontologies, for describing functional genomics data [1] and have found that the area of phenotype description remains a major challenge. Within this the formal description of diseases and disease processes, in a way that can be used across species, is a vital but currently unmet need. It is this problematical area that we review in this paper.

## II. Coding of phenotype data

Much phenotype data is traditionally described in natural language, frequently using a mixture of unstructured terminologies or free text, often with variations in practice within a single knowledge domain. Quantitative data are presented using disparate data models and indexed with simple text descriptions, either local controlled vocabularies, or, at best, with terms from UMLS or MeSH. The reason for this is that natural language is highly expressive and the range of information captured in phenotype descriptions is usually both deep and broad; hence natural language is the most obvious medium in which to record and express it. However natural language is hard to compute on and suffers

Manuscript received 25th July 2008. This work was funded by the Commission of the European Community Contract number LSHG-CT-2006-037811. JPS acknowledges support of the US National Institutes of Health (CA089713) and the Ellison Medical Foundation.

This paper is dedicated to the memory of Victor McKusick whose visionary approach shaped medical genetics in the 20th and 21st centuries. *Pigmaei gigantum humeris impositi plusquam ipsi gigantes vident*

P. N. Schofield is with the Department of Physiology Development and Neuroscience at the University of Cambridge, UK; (phone: 44-1223-333878; fax: 44-1223-333840.; email: ps@mole.bio.cam.ac.uk.)

G. V. Gkoutos is with the Department of Genetics, University of Cambridge, Downing Street, Cambridge and Institute of Biological Research and Biotechnology, National Hellenic Research Foundation, Greece; (email: gg295@cam.ac.uk)

J. Sundberg is with The Jackson Laboratory, Bar Harbor, Maine, USA; (email: JPS@jax.org)

J. M. Hancock is with the Bioinformatics Group, MRC Harwell, Harwell, Oxfordshire, UK; (email: j.hancock@har.mrc.ac.uk)

CASIMIR Partners: University of Cambridge, Cambridge, UK; MRC Harwell, Oxfordshire, UK; MRC, Edinburgh, UK; EBI, Hinxton, UK; EMBL, Monterotondo, Italy; BSRC Fleming, Vari, Greece; GSF National Research Center for Environment and Health, Neuherberg, Germany; Helmholtz-Zentrum fuer infektionsforschung GmbH, Braunschweig, Germany; CNR-Consiglio Nazionale delle Richerche-Instituto di Biologia Cellulare, Monterotondo, Italy; Geneservice Limited, Cambridge, UK

from the now often rehearsed problems of ambiguity, semantic complexity and lack of structure.

Within the mouse community there are currently two types of database carrying phenotype related information which have an emphasis on either qualitative or quantitative phenotype descriptions, although these are not mutually exclusive [4].

Qualitative data such as that held in the Mouse Genome Informatics databases (MGI) [5], [6] is mainly coded by the mammalian phenotype ontology (MP) [7]. This is currently the most successful and readily applicable approach to describing a wide range of aspects of phenotype and disease in a set of carefully defined descriptive terms, variously capturing abnormal traits, abnormal processes, summative diagnoses and other descriptors of phenodeviance; deviance of a phenotype, such as weight, coat colour, blood metabolites etc. in the population under investigation from that in the reference population. The upper level terms of the MP ontology include physiological systems, behaviour, developmental phenotypes and aging and below, this level physiological systems are divided into morphological and physiological phenotypes. Much manifest disease can be coded readily by MP and currently there are 88,600 annotations of approximately 21,000 genotypes. This is a "classically structured" directed acyclic graph based ontology and is designed in such a way as to enable searching of phenotype databases to find mutations and alleles with specific phenotypes, allow gene clustering based on mutant phenotypes, the discovery of genes in related pathways or potential mouse models of human diseases. These applications neatly summarise the utility of formal frameworks for disease description. However there are problems for cross-species phenotype matching, with this approach, both in the range and type of terms used and in the organization which as yet does not allow the ontology to be used for inference. That there is currently no common vocabulary to describe phenotypes across different organisms is a major challenge, as discussed below.

The other class of phenotype information is based on quantitative data, mainly that for quantitative and complex traits. Three major databases exemplify this type of approach, the Mouse Phenome Database [8], Europhenome [9] and Gene Network (including WebQTL) [10]. The former two concentrate on quantitative data on traits in background strains of mice. Gene Network contains data from reference populations of mice, rats and *Arabidopsis* for a wide range of complex traits such as cancer susceptibility, toxicity, and behaviour. Europhenome will in the future hold phenotype data on mutant strains, particularly those derived from the global mouse knockout efforts currently underway [11] not all of which will be quantitative. For a fuller description of mouse phenome resources see Hancock & Mallon [4].

For human data the situation is much more complex. The call for a human Phenome Project in 2003 [12] with emphasis on the need for standards and international integration has not yet met with a concerted response, and it is still fair to say that with regard to human phenotypes and traits there is an un-coordinated scatter of data throughout databases and resources across the world. Much human phenotype data relates to disease and its predisposition, and is largely captured with free text. In the best situations it may be coded using clinical informatics formalisms such as ICD9/10, SNOMED. Although quantitative trait and complex trait data is also available, this varies in structure and utility.

Human genetic databases may be divided into core databases and locus specific databases (LSDB) where there is either an attempt to provide data on all pathological variation and its consequences, such as the Human Gene Mutation Database HGMD [13] which uses a local controlled vocabulary, or only on one gene or locus respectively (see Patrinos and Brookes for discussion [14]). The genetic association database (GAD) [15], for example, contains associations between complex diseases and disorders and individual human genes curated from the literature; here diseases are categorised using a controlled vocabulary drawn from MeSH terms. Quantitative data sets on human populations are held by DBGaP [16] and again indexed in a largely unstructured way through MeSH defined terms. LSDB databases are usually manually curated and contain unpublished information which includes genetic variation data not currently associated with an abnormal phenotype or pathology.

Both classes of database describe the phenotypic consequences of the mutations they contain, most often in free text or occasionally in a mixture of clinical informatics terms and natural language. The most encyclopaedic resource is of course Online Mendelian Inheritance in Man (OMIM) [17], but this exemplifies the problems with human disease databases in that the disease descriptions are not only in natural language, but each record is historically cumulative, so although this is the gold standard resource for researching the effects of single Mendelian alleles automated use of OMIM data, for example by text mining, can lead to serious errors in assigning disease terms to genes.

A classic example of how lack of easy access to the literature combined with inaccurate curation led to an error that eventually led to a much greater understanding of a particular disease involved the hairless gene and its incorrect link to the complex polygenic disease known as alopecia universalis. The hairless phenotype and its more severe form, known as rhino (short for rhinoceros), was first described in mice in 1856 [18]. The human homolog, atrichia with papules or as it later became known as, papular atrichia, was first described in 1954, nearly 100 years later [19]. The correlation between the mouse and human disease was made some 30 years afterwards [20], [21] The hairless gene was linked to a simple, recessively inherited form of alopecia universalis based on the Online Mendelian Inheritance in Man entry (OMIM: 203655) [22]. The OMIM designation was based on morphologic diagnosis; total lack of hair in patients with an autosomal recessive pattern of inheritance. Alopecia universalis is actually a well-characterized complex genetic based autoimmune skin

disease in both humans [23] and mice [24]. While this mismatch was initially of great concern [25] it subsequently led to a much better understanding of papular atrichia. Many mutations have now been identified in the human hairless gene as well as in rodents and non-human primates [26], [27].

## III. Crossing the species divide; granularity and specificity

We now have many examples of the power of using phenotype descriptions to discover new relationships between genes and phenotypes and new functions for previously uncharacterised genes and alleles. A good example is PhenomicDB [28] which contains one of the most wide ranging cross-species datasets on gene/phenotype associations through combining data from OMIM, the Mouse Genome Database (MGD), WormBase, FlyBase, the Comprehensive Yeast Genome Database (CYGD), the Zebrafish Information Network (ZFIN), and the MIPS *Arabidopsis thaliana* database (MAtDB). Groth et al [29] queried the resulting "warehouse" using a text-mining approach which generated a summary phenotypic statement for each gene, then clustered the statements to produce what Brunner et al [30] have termed "Phenoclusters" – a group of genes with overlapping phenotypes, which may then be used for discovery of new disease or functional associations. This phenotype driven approach to discovery of gene function has distinct advantages to the gene driven approach to phenotype prediction as whilst many closely related phenotypes are caused by mutations in different genes whose gene products interact directly or are on the same pathway, mutations in the same gene can have diverse phenotypic outcomes depending on which function of a multifunctional gene product are compromised. Several related disease candidate gene discovery approaches have been developed (see Tiffin et al. [31] and Oti and Brunner [30] and van Driel et al. [32] for review) but in the absence of systematic coding all depend to a greater or lesser extent on text-mining from their data sources, and making use, at best, of UMLS and MeSH terms in abstracts and database phenotype fields Despite impressive results from many of these approaches it is clear that a standardised description of the phenotype or disease would greatly increase their power and specificity.

A key issue is the assumption that the currently dominant paradigm for disease conceptualization is useful for all applications. It is a mistake to assume that the human "phenome" is a list of "diseases" which form more or less distinct entities. The realization that diseases of separate genetic aetiology may share similar phenotypes may seem obvious, but it is only recently that this has generated attention. Work by Brunner and others [30], [33] demonstrated that shared aspects of phenotype may be viewed as a proxy for a common underlying pathogenetic mechanism and that this mechanism may be shared by dysfunction of a group of genes whose products either interact or are on the same functional pathway. This "modularity" of phenotypes should not come as a surprise, but it makes the formulation of a new concept of disease description all the more urgent. The generation of phenoclusters depends on the ability to code phenotypes in as granular way as possible. This approach was originally used in making gene/phenotype associations in RNAi generated phenotypes in *C. elegans* where each phenotype was expressed as a combination of 45 phenotypic features, enabling clustering of functionally related genes [34].

Use of a phenotype driven approach to discover new information about gene/phenotype relationships *within* a species requires a sufficiently high level of specificity and granularity to be able to discriminate between closely related phenotypes with overlapping components. This is particularly true of complex traits. Joy and Hegele [35] provide an excellent discussion of the problems caused by the accuracy and variability of definitions in the context of Metabolic Syndrome and the resulting problems with candidate gene association and linkage studies; similar problems dogged gene association studies in X-linked mental retardation where there are insufficient phenotypic features to "unbundle" non-syndromic cases in gene association studies [36].

The requirement for "deep phenotyping" using well defined criteria is therefore clearly of importance in human gene association studies, but it is also crucial if human phenotypes are to be compared to those from model organisms. The deficiency is well demonstrated by the analysis of cross-species phenoclustering carried out using PhenomicDB by Groth et al. discussed above. More than 90% of the clusters they generated contained genes from a single species and there was a tendency of genes to fall into species specific clusters. They interpret this to indicate that the terminology used to describe phenotype in each species fails to cross the species barrier even though many phenotypes clearly have their equivalents between species. It is therefore clear that we need a change in the way in which we describe disease if our aim is to understand the underlying processes and genetic aetiology through using model organisms.

The principles of "One Medicine" originate with Sir William Osler (1849-1919), and were affirmed by Rudolf Virchow (1821-1902) in his contention that both animal and human pathology had common underlying principles and manifestations.

*"Between animal and human medicine there is no dividing line - nor should there be. The object is different but the experience obtained constitutes the basis of all medicine."* [37]

Indeed, the assumption of "One Medicine" is the fundamental underpinning of the use of model organisms to study human disease [38], [39]. With this principle accepted it should be possible to derive a description framework which crosses between species to capture common pathological processes and outcomes.

Current disease description frameworks are designed for a particular purpose. Medical informatics and the tools derived primarily from medical-informatic approaches are constructed to deal with the practical issues of record

keeping, billing and communication between professionals. This level of abstraction is pragmatic and powerful for the purposes for which it was intended. A consequence of this is that formalisms based on a clinical model, such as ICD9, are potentially very useful for text mining from the human clinical literature as they deal with shared abstractions and "concepts" and to a large degree benefit from historical consistency (with some notable exceptions). However, for reasons described above they do not allow inference in a biologically meaningful way, partly for intrinsic structural reasons, but also because so much of this framework is concerned with diagnosis and larger "disease concepts" which wrap and obscure the individual manifestations of the disease which give clues to its pathobiology. Being largely composed of summative terms they do not readily allow for the investigation of shared aspects of diseases or phenotypes which were hitherto unknown.

TABLE I

| MP | DO | DO |
|---|---|---|
| | HELLP syndrome: DO:0013133 | |
| ?Pregnancy related premature death. MP 0008028? | | Disorders of pregnancy: DO:0005366 |
| Hypertension. MP:0001595 | | Hypertension. DO:00010763 |
| | | Hypertension associated with pregnancy: DO: 0005365 |
| Thrombocytopenia. MP:0003179 | | Thrombocytopenia. DO:0001588 |
| Renal Failure. MP: 0003606 | | Renal failure. DO: 0001074 |
| Hepatic failure. MP 0003326 | | Liver failure. DO: 0004722 |
| Hepatic necrosis. MP: 0001654 | | Acute and subacute liver necrosis. DOID:0014551 |
| Proteinuria: MP. 0005160 | | Proteinuria. DO: 0000576 |
| Abnormal glomeruli. MP: 0005160 | | Glomerular Vascular Disorder: DO:0002976 |
| Abnormal labyrinth layer. MP: 0001716 | | ?Placenta disorder: DO: 0000780? |
| Haemolytic anaemia. MP:0001585 | | Anemia haemolytic. DO: 0000583 |
| | | Disseminated Intravascular Coagulation. DO: 0011247 |

Logical decomposition of HELLP using Disease Ontology (DO) and closest matching terms in the Mammalian Phenotype (MP) ontology. Effective identity of endophenotypes is found for hypertension, thrombocytopenia, renal failure, hepatic failure, hepatic necrosis, proteinuria and haemolytic anaemia. Approximate, pragmatically useful matches are found with pregnancy related premature death, abnormal glomeruli, abnormal labyrinth layer, but none for DIC. The best matches are found unsurprisingly in the common pathological descriptions and the poorest where there is indication of aetiology, e.g. "pregnancy related".

An example of this problem is shown in Table I using the Disease Ontology (DO; http://diseaseontology.sourceforge.net/) a large directed acyclic graph based on UMLS constituent vocabularies which is currently the most complete human disease ontology available. Here we have taken the HELLP syndrome ("H" haemolysis "EL" for elevated liver enzymes, and "LP" for low platelet count), an acute complication of pre-eclampsia [40]. Both HELLP syndrome and pre-eclampsia are contained within the Disease Ontology (v3) and within ICD9, but neither feature in the MP mammalian phenotype ontology which is used to code mouse mutants and strains. Both are "disease concepts" in that they encapsulate related diagnostic entities composed of a set of underlying lesions and pathological manifestations. The entities into which complex "concepts" may be decomposed or disaggregated may be termed "endophenotypes". This is a concept originally developed in psychogenetics [41], [42]. As long ago as 1973 Gottesman et al [43] commented that it had become clear that "*classification of psychiatric diseases on the basis of overt phenotypes (syndromic behaviors) might not be optimal for genetic dissection of these diseases, which have complex*

*genetic underpinnings"* and, borrowing from the *Drosophila* genetics world [44], the term "endophenotype" was adopted to describe a constituent of a complex phenotype into which the diagnosis could be decomposed or deconstructed to facilitate genetic analysis. Searching of both MP and DO reveals a set of terms which constitute a spectrum of the known endophenotypes of HELLP syndrome, with a significant degree of overlap. It is interesting that the best matches are found for the most basic pathological lesions and processes, common to both organisms, and the worst for the aetiologically predicated terms, such as those referring to "pregnancy related…" aspects. It is clear that coding the disease concept with a single term will not allow discovery of related disease phenotypes in other species, in this case the mouse, while coding with a combination of endophenotypes stand a much better chance of discovering a cross-species phenocluster.

## IV. Decomposition and logical definition

Coding of phenotype data is a laborious expert task, especially from the literature, and consequently it is not feasible to expect curators to code disease data manually using endophenotypes alone, indeed the expertise required to break down "disease concepts" into endophenotypes is considerable. Loss of the "disease concept " term, such as "HELLP", in high resolution coding is highly undesirable. We suggest that the way forward is to provide endophenotype coding as a logical definition of each higher level concept based on the approach implemented elsewhere [45]. Such endophenotypes, being closer to the underlying pathobiology, are likely to be shared between different organisms as they reflect much more basic processes and responses to underlying lesions. This brings us directly back to Virchow's assertion of the unity of animals and man through shared pathological processes which at least within vertebrates show remarkable evolutionary conservation.

It is clear that disease concepts can be broken down into endophenotypes using terms for constituent processes in the Disease Ontology, but this may also be done with other ontologies, for example MPATH the mouse pathology ontology [46]. The MoDIS database capture tool, currently in use at the Jackson Laboratory [47] allows the working pathologist to code diagnoses using a combination of ontologies, MPATH and MA, the adult mouse anatomy ontology [48] and to relate these to higher order disease concepts or summative diagnoses. Examples of such disaggregated codings are shown in Table II.

We propose that disaggregation of disease concepts into their constituent endophenotypes has the potential to form a bridge between disease related phenotypes of human, mouse and other vertebrates, as they constitute a common vocabulary of pathological response to underlying lesions and in combination characterize the emergent phenotype.

## V. Phenotypic qualities and a combinatorial approach to logical definitions

We have suggested above how disaggregation of "disease concepts" into the underlying endophenotypes will allow easier cross-species matching of otherwise species specific disease entities which share a common pathobiology. The resulting endophenotypes are in all cases amenable to further logical definition based on the measurement of variables such as size, frequency, behaviour, alteration of metabolite levels etc. these objective measurements underlie both model organism phenotyping and in recent years the concept of "Evidence-based medicine" [49] where the aim is to make *"conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients."* [50] based on objective application of the scientific method in determining best practice. We have proposed elsewhere an approach based on a so-called decomposition methodology [45] that makes use of a combination of relevant descriptive ontologies, such as MP, CheBI [51] etc. so that the deconstructed terms are expressed in an Entity+Quality (E+Q) formalism [52] An application example that uses this approach for the mouse is cited in the paper from Beck and co workers [53].

Phenotype terms could be deconstructed employing terms to describe the affected bearer entities from various core ontologies and appropriate quality terms from an ontology of qualities termed PATO [51]. Following this methodology the MP term *belly spot* (MP:0000373) can be decomposed to *spotted* (PATO:0000333) [has quality] *white* (PATO:0000323) [inheres_in] *coat hair* (MA:0000155) [part of] *abdomen* (MA:0000029).

Both concept (in this case "belly spot") and deconstructed terms are stored in the Europhenome database for similar reasons as discussed above for the endophenotype and disease concept in the context of disease coding.

In this approach a distinction is made between qualitative and quantitative phenotype data, as the annotation of these two classes of data presents different problems and advantages. Qualitative measurements are scored on the basis on the judgment of the individual investigator as to whether a particular individual is abnormal in comparison to the reference. This is unsatisfactory because it relies on individual judgment and because it is not readily interpretable by the non-expert. Ideally such quantitative measures will be gradually replaced by more objective measures where calling the phenotype term depends on statistical calculations and pre-determined conditions.

## VI. Conclusions

The work carried out by CASIMIR to date suggests that ontologies and description frameworks for capturing data on disease and phenotype are essential tools for the support of mouse functional genomics, and in a broader context for the assignation of functions to genes. At the moment, although tools are available, they are still in the early stages of development and may need to be applied in new ways to fully serve the requirements of cross species phenotype mapping. Even a preliminary attempt to implement existing ontologies in the E+Q framework demonstrates the need for more terms to describe measured entities, both in humans

TABLE II

| Organ (MA) | Pathology ID (MPATH) | Definitive Diagnosis | Disease Ontology ID (DO) | Reference |
|---|---|---|---|---|
| Skin MA:0000151 | Immune mediated disease MPATH: 0000194<br>Alopecia MPATH: 0000025 | Alopecia areata | Alopecia areata DO: 0000986 | King LE, McElwee KJ, Sundberg JP: Alopecia areata. Edited by Nickoloff BJ, Nestle FO. Basel, Karger, 2008, p.pp. 280-312<br>McElwee K, Boggess D, Miller J, King L, Sundberg J: Spontaneous alopecia areata-like hair loss in one congenic and seven inbred laboratory mouse strains, J Invest Dermatol Symp Proc 1999, 4 202-206 |
| Skin MA:0000151 | Granulation tissue MPATH: 0000183<br>Fluid accumulation MPATH: 0000106 | Pemphigus vulgaris | Pemphigus vulgaris. DO: 00009182 | Montagutelli X, Lalouette A, Boulouis HG, Guenet J-L, Sundberg JP: Vesicle formation and follicular root sheath separation in mice homozygous for deleterious alleles at the balding (bal) locus, J Invest Dermatol 1997, 109:324-328<br>Sundberg JP: The balding (bal) mutation, chromosome 18. Edited by Sundberg JP. Boca Raton, CRC Press, 1994, p.pp. 187-191<br>Sundberg JP, Shultz LD, King LE, Montagutelli X: The spontaneous balding and desmoglein 3 null mutations: mouse models for pemphigus vulgaris, Comp Pathol Bull 1998, 30:3-4 |
| Vibrissa MA:0000163 | Calcification MPATH: 0000036<br>Tissue specific degenerative process MPATH: 0000025<br>Extracellular and intracellular depletion MPATH: 0000047 | Pseudoxanthoma elasticum | Pseudoxanthoma elasticum. DO: 0002738 | Klement JF, Matsuzaki Y, Jiang QJ, Terlizzi J, Fujimoto N, Li K, Pulkkinen L, Birk DE, Sundberg JP, Uitto J: Targeted ablation of the Abcc6 gene results in ectopic mineralization of connective tissues, Mol Cell Biol 2005, 2005 8299-8310 |
| Caecum MA: 0000334 | Acute inflammation MPATH: 0000213<br>Ulcerative inflammation MPATH: 0000217 | Inflammatory bowel disease | | Sundberg JP, Elson CO, Bedigian H, Birkenmeier EH: Spontaneous, heritable colitis in a new substrain of C3H/HeJ mice, Gastroenterology 1994, 107:1726-1735<br>Bristol IJ, Farmer MA, Cong Y, Zheng XX, Strom TB, Elson CO, Sundberg JP, Leiter EH: Heritable susceptibility for colitis in mice induced by IL-10 deficiency, Inflamm Bowel Dis 2000, 6 290-302 |
| Skeletal Muscle MA:0000165 | Rhabdomyosarcoma MPATH: 0000428 | Rhabdomyosarcoma | Rhabdomyosarcoma. DO: 0003247 | Sundberg JP, Adkison DL, Bedigian HG: Skeletal muscle rhabdomyosarcomas in inbred laboratory mice, Vet Pathol 1991, 28:200-206 |
| Eyelid MA:0000168 | Acute inflammation MPATH: 0000213<br>Ulcerative inflammation. MPATH: 0000217 | Ulcerative blepharitis | Ulcerative blepharitis. DO: 0009483 | Sundberg JP, Brown KS, Bedigian R: Ulcerative blepharitis and periorbital abscesses in BALB/cJ and BALB/cByJ mice, JAX notes 1990, 443 3-4<br>Smith RS, Montagutelli X, Sundberg JP: Ulcerative blepharitis in aging inbred mice. Edited by Mohr U, Dungworth DL, Capen CC, Carlton W, Sundberg J, Ward J. Washington, D.C., ILSI Press, 1996, p.pp. 131-138 |
| Lens MA:0000275 | Cataract; nuclear and cortical MPATH: 0000462 | Nuclear cataract | Nuclear senile cataract. DO: 0013963 | Smith RS, Sundberg JP, Linder CC: Mouse mutations as models for studying cataracts, Pathobiology 1997, 65:146-154 |
| Bone MA:0001459 | Hyperplasia MPATH: 0000134<br>Fibro-osseous lesion MPATH: 0000590 | Fibro-osseous lesion | Possibly fibro-osseous digital pseudotumor. DO: 0008153 ? | Abbassan MA, Wojcinski ZW, Barsoum NJ, Smith, GS (1991) Spontaneous fibro-osseous proliferative lesions in the sternums and femurs of B6 C3Fl mice. Vet Path 28: 381-388 |

Decomposition of definitive diagnoses and disease concepts into a combination of MA and MPATH ontology terms and matching with Disease ontology concepts. Good matches are found for pemphigus vulgaris, pseudoxanthoderma

elasticum, rhabdomyosarcoma, ulcerative blepharitis provide good matches, but these are less good with nuclear cataract and fibroosseous lesions. There is currently no term for IBD in DO.

and in mice, and for example a mammalian trait ontology would be of great utility. With respect to the human it will not always be possible to obtain or record measurements with the same completeness or precision as with mice in a laboratory setting, and in many cases phenotype description from the literature will inevitably be only qualitative, if only because it constitutes legacy data. The power of the decompositional approach is that it may be applied in both qualitative and quantitative manners and in either lends itself to computational analysis. The amount of work which needs to be done is daunting, but the realisation of the scale and importance of the work should encourage funding agencies to prioritise such an effort.

## REFERENCES


[1] J. M. Hancock, P. N. schofield, C. Chandras, M. Zouberakis, V. Aidinis, D. Smedley, N. Rosenthal, K. Schughart, and The_CASIMIR_Consortium, "CASIMIR: Coordination and Sustainability of International Mouse Informatics Resources," *8th IEEE International Conference on Bioinformatics and Bioengineering*, vol. To be published, 2008.

[2] M. Zouberakis, C. Chandras, J. M. Hancock, P. N. Schofield, and V. Aidinis, "The Mouse Resource Browser (MRB) – A near-complete registry of mouse resources," *8th IEEE International Conference on Bioinformatics and Bioengineering,*, vol. To Be Published, 2008.

[3] C. Chandras, T. Weaver, M. Zouberakis, J. M. Hancock, P. N. schofield, and V. Aidinis, "Digital Preservation – Financial Sustainability of Biological Data and Material Resources," *8th IEEE International Conference on Bioinformatics and Bioengineering*, vol. To be published, 2008.

[4] J. M. Hancock and A. M. Mallon, "Phenobabelomics--mouse phenotype data resources," *Brief Funct Genomic Proteomic*, vol. 6, pp. 292-301, Dec 2007.

[5] J. T. Eppig, J. A. Blake, C. J. Bult, J. E. Richardson, J. A. Kadin, and M. Ringwald, "Mouse genome informatics (MGI) resources for pathology and toxicology," *Toxicol Pathol*, vol. 35, pp. 456-7, 2007.

[6] C. J. Bult, J. T. Eppig, J. A. Kadin, J. E. Richardson, and J. A. Blake, "The Mouse Genome Database (MGD): mouse biology and model systems," *Nucleic Acids Res*, vol. 36, pp. D724-8, Jan 2008.

[7] C. L. Smith, C. A. Goldsmith, and J. T. Eppig, "The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information," *Genome Biol*, vol. 6, p. R7, 2005.

[8] M. A. Bogue, S. C. Grubb, T. P. Maddatu, and C. J. Bult, "Mouse Phenome Database (MPD)," *Nucleic Acids Res*, vol. 35, pp. D643-9, Jan 2007.

[9] A. M. Mallon, A. Blake, and J. M. Hancock, "EuroPhenome and EMPReSS: online mouse phenotyping resource," *Nucleic Acids Res*, vol. 36, pp. D715-8, Jan 2008.

[10] J. Wang, R. W. Williams, and K. F. Manly, "WebQTL: web-based complex trait analysis," *Neuroinformatics*, vol. 1, pp. 299-308, 2003.

[11] F. S. Collins, J. Rossant, and W. Wurst, "A mouse for all reasons," *Cell*, vol. 128, pp. 9-13, Jan 12 2007.

[12] N. Freimer and C. Sabatti, "The human phenome project," *Nat Genet*, vol. 34, pp. 15-21, May 2003.

[13] P. D. Stenson, E. Ball, K. Howells, A. Phillips, M. Mort, and D. N. Cooper, "Human Gene Mutation Database: towards a comprehensive central mutation database," *J Med Genet*, vol. 45, pp. 124-6, Feb 2008.

[14] G. P. Patrinos and A. J. Brookes, "DNA, diseases and databases: disastrously deficient," *Trends Genet*, vol. 21, pp. 333-8, Jun 2005.

[15] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, "The genetic association database," *Nat Genet*, vol. 36, pp. 431-2, May 2004.

[16] M. D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, N. Popova, S. Pretel, L. Ziyabari, M. Lee, Y. Shao, Z. Y. Wang, K. Sirotkin, M. Ward, M. Kholodov, K. Zbicz, J. Beck, M. Kimelman, S. Shevelev, D. Preuss, E. Yaschenko, A. Graeff, J. Ostell, and S. T. Sherry, "The NCBI dbGaP database of genotypes and phenotypes," *Nat Genet*, vol. 39, pp. 1181-6, Oct 2007.

[17] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res*, vol. 33, pp. D514-7, Jan 1 2005.

[18] J. S. Gaskoin, "On a peculiar variety of Mus musculus," *Procedings of the Zoological Society of London*, vol. 24, pp. 38-40, 1856.

[19] J. Damste and J. R. Prakken, "Atrichia with papular lesions: a variant of congenital ectodermal dysplasia," *Dermatologica*, vol. 108, pp. 114-121, 1954.

[20] J. P. Sundberg, R. W. Dunstan, and J. G. Compton, *Hairless mouse, HRS/J hr/hr*. Heidelberg: Springer-Verlag, 1989.

[21] J. P. Sundberg, "The hairless (hr) and rhino (hrrh) mutations, chromosome 14," in *Handbook of mouse mutations with skin and hair abnormalities: animal models and biomedical tools*, J. P. Sundberg, Ed. Boca Raton: CRC Press, 1994, pp. 291-312.

[22] W. Ahmad, U. Faiyaz, V. Brancolini, H. C. Tsou, S. u. Haque, H. Lam, V. M. Alta, J. Owen, M. deBlaquiere, J. Frank, P. B. Cserhalmi-Friedman, A. Leask, J. A. McGrath, M. Peacocke, M. Ahmad, J. Ott, and A. M. Christiano, "Alopecia universalis associated with a mutation in the human hairless gene," *Science*, vol. 279, pp. 720-724, 1998.

[23] A. Martinez-Mir, A. Zlotogorski, D. Gordon, L. Petukhova, J. Mo, T. C. Gilliam, D. Londono, C. Haynes, J. Ott, M. Hordinsky, K. Nanova, D. Norris, V. Price, M. Duvic, and A. M. Christiano, "Genomewide scan for linkage reveals evidence of several susceptibility loci for alopecia areata," *Am J Hum Genet*, vol. 80, pp. 316-328, 2007.

[24] J. P. Sundberg, K. A. Silva, R. Li, L. E. King, and G. A. Cox, "Adult onset alopecia areata is a complex polygenic trait in the C3H/HeJ mouse model," *J Invest Dermatol*, vol. 123, pp. 294-297, 2004.

[25] J. P. Sundberg, V. H. Price, and L. E. King, "The "hairless" gene in mouse and man," *Arch Dermatol*, vol. 135, pp. 718-720, 1999.

[26] W. Ahmad, M. S. Ratteree, A. A. Panteleyev, V. M. Aita, J. P. Sundberg, and A. M. Christiano, "Atrichia with papular lesions resulting from mutations in the rhesus macaque (Macaca mulatta) hairless gene," *Lab Anim*, vol. 36, pp. 61-67, 2002.

[27] A. A. Panteleyev, R. Paus, W. Ahmad, J. P. Sundberg, and A. M. Christiano, "Molecular and functional aspects of the hairless (hr) gene in laboratory rodents and humans," *Exp Dermatol*, vol. 7, pp. 249-267, 1998.

[28] P. Groth, N. Pavlova, I. Kalev, S. Tonov, G. Georgiev, H. D. Pohlenz, and B. Weiss, "PhenomicDB: a new cross-species genotype/phenotype resource," *Nucleic Acids Res*, vol. 35, pp. D696-9, Jan 2007.

[29] P. Groth, B. Weiss, H. D. Pohlenz, and U. Leser, "Mining phenotypes for gene function prediction," *BMC Bioinformatics*, vol. 9, p. 136, 2008.

[30] M. Oti and H. G. Brunner, "The modular nature of genetic diseases," *Clin Genet*, vol. 71, pp. 1-11, Jan 2007.

[31] N. Tiffin, E. Adie, F. Turner, H. G. Brunner, M. A. van Driel, M. Oti, N. Lopez-Bigas, C. Ouzounis, C. Perez-Iratxeta, M. A. Andrade-Navarro, A. Adeyemo, M. E. Patti, C. A. Semple, and W. Hide, "Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes," *Nucleic Acids Res*, vol. 34, pp. 3067-81, 2006.

[32] M. A. van Driel and H. G. Brunner, "Bioinformatics methods for identifying candidate disease genes," *Hum Genomics*, vol. 2, pp. 429-32, Jun 2006.

[33] M. Oti, M. A. Huynen, and H. G. Brunner, "Phenome connections," *Trends Genet*, vol. 24, pp. 103-6, Mar 2008.

[34] F. Piano, A. J. Schetter, D. G. Morton, K. C. Gunsalus, V. Reinke, S. K. Kim, and K. J. Kemphues, "Gene clustering based on RNAi phenotypes of ovary-enriched genes in C. elegans," *Curr Biol*, vol. 12, pp. 1959-64, Nov 19 2002.

[35] T. Joy and R. A. Hegele, "Genetics of metabolic syndrome: is there a

role for phenomics?," *Curr Atheroscler Rep*, vol. 10, pp. 201-8, Jun 2008.
[36] H. H. Ropers and B. C. Hamel, "X-linked mental retardation," *Nat Rev Genet*, vol. 6, pp. 46-57, Jan 2005.
[37] R. Virchow, *Cellular pathology as based upon physiological and pathological histology*. New York: R M DeWitt, 1860.
[38] N. Rosenthal and S. Brown, "The mouse ascending: perspectives for human-disease models," *Nat Cell Biol*, vol. 9, pp. 993-999, 2007.
[39] Committee_on_the_National_Needs_for_Research_on_Agriculture_and_Natural_Resources_National_Research_Council_of_the_National_Academies, *Critical needs for research in veterinary science*. Washington DC: The National Academies Press, 2005.
[40] D. Mihu, N. Costin, C. M. Mihu, A. Seicean, and R. Ciortea, "HELLP syndrome - a multisystemic disorder," *J Gastrointestin Liver Dis*, vol. 16, pp. 419-24, Dec 2007.
[41] Gottesman, II and T. D. Gould, "The endophenotype concept in psychiatry: etymology and strategic intentions," *Am J Psychiatry*, vol. 160, pp. 636-45, Apr 2003.
[42] T. D. Gould and Gottesman, II, "Psychiatric endophenotypes and the development of valid animal models," *Genes Brain Behav*, vol. 5, pp. 113-9, Mar 2006.
[43] Gottesman, II and J. Shields, "Genetic theorizing and schizophrenia," *Br J Psychiatry*, vol. 122, pp. 15-30, Jan 1973.
[44] B. John and K. R. Lewis, "Chromosome Variability and Geographic Distribution in Insects," *Science*, vol. 152, pp. 711-721, May 6 1966.
[45] C. M. Mungall and G. V. Gkoutos, "Reconciling Phenotype descriptions across multiple species," in *Submitted*, 2008.
[46] P. N. Schofield, J. B. Bard, C. Booth, J. Boniver, V. Covelli, P. Delvenne, M. Ellender, W. Engstrom, W. Goessner, M. Gruenberger, H. Hoefler, J. Hopewell, M. Mancuso, C. Mothersill, C. S. Potten, L. Quintanilla-Fend, B. Rozell, H. Sariola, J. P. Sundberg, and A. Ward, "Pathbase: a database of mutant mouse pathology," *Nucleic Acids Res*, vol. 32, pp. D512-5, Jan 1 2004.
[47] J. P. Sundberg, Sundberg, B. , Schofield, P.N., "Integrating Mouse Anatomy and Pathology Ontologies into a Phenotyping Database: Tools for Data Capture and Training," *Mamm Genome*, 2008.
[48] T. F. Hayamizu, M. Mangan, J. P. Corradi, J. A. Kadin, and M. Ringwald, "The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data," *Genome Biol*, vol. 6, p. R29, 2005.
[49] D. L. Sackett and W. M. Rosenberg, "The need for evidence-based medicine," *J R Soc Med*, vol. 88, pp. 620-4, Nov 1995.
[50] D. L. Sackett, W. M. Rosenberg, J. A. Gray, R. B. Haynes, and W. S. Richardson, "Evidence based medicine: what it is and what it isn't," *Br Med. J.*, vol. 312, pp. 71-2, 1996.
[51] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, and M. Ashburner, "ChEBI: a database and ontology for chemical entities of biological interest," *Nucleic Acids Res*, vol. 36, pp. D344-50, Jan 2008.
[52] C. M. Mungall, G. V. Gkoutos, S. E. Washington, and S. E. Lewis, "Representing Phenotypes in OWL," in *Proceedings of the OWLED workshop on OWL*, 2007.
[53] T. Beck, A. M. Mallon, H. Morgan, A. Blake, and J. M. Hancock, "Using ontologies to annotate large-scale mouse phenotype data," in *Proceedings 11th Annual Bio-Ontologies Meeting, ISMB*, Toronto, 2008.